# Einführung in die Verlaufsdatenanalyse: statistische Grundlagen und Anwendungsbeispiele zur Längsschnittanalyse kategorialer Daten

Andreß, Hans-Jürgen

Veröffentlichungsversion / Published Version
Themenheft / topical issue

**Empfohlene Zitierung / Suggested Citation:**
Andreß, H.-J. (1992). Einführung in die Verlaufsdatenanalyse: statistische Grundlagen und Anwendungsbeispiele zur Längsschnittanalyse kategorialer Daten. *Historical Social Research, Supplement*, 5, 1-323. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-285968

HSR Supplement/Beiheft No. 5 (1992)

Hans-Jürgen Andreß

# Einführung in die Verlaufsdatenanalyse

Statistische Grundlagen und Anwendungsbeispiele zur Längsschnittanalyse kategorialer Daten

Köln
Zentrum für Historische Sozialforschung
1992

*Hans-Jürgen Andreß*

# Einführung in die Verlaufsdatenanalyse

## Statistische Grundlagen und Anwendungsbeispiele zur Längsschnittanalyse kategorialer Daten

### *Inhaltsverzeichnis*

# *Vorwort*

Im Herbst 1991 anläßlich des ZHSF–Herbstseminars über "Event History Analysis" bot sich mir die Gelegenheit, einen Text zu überarbeiten, der 1985 als Band 1 der ZUMA–Methodentexte erschien. Das Ergebnis liegt nunmehr vor und ich möchte es Günter Albrecht, Herwig Birg, Theodor Harder und Peter Naeve widmen, die mich durch ihre Kritik zu dieser Überarbeitung animiert haben – nicht zu vergessen Manfred Küchler, der das ganze Unternehmen von Anfang an gefördert hat. Zur Dank verpflichtet bin ich natürlich auch den Teilnehmern und Teilnehmerinnen verschiedener Seminare, die mich durch ihre Fragen, Kritik und Verbesserungsvorschläge gezwungen haben, die Darstellung der Materie zu präzisieren und verständlicher zu machen. Angesichts der mathematisch–statistischen Voraussetzungen war dieses kein einfaches Unterfangen.

Auf dem Wege der Überarbeitung hat das Werk seine soziologische Herkunft nicht ganz abstreifen können. Ich hoffe, dies ist für Historiker und Historikerinnen kein zu großes Hindernis. Ich habe mir jedoch sagen lassen, daß die Analyse von Lebens– und Berufsverläufen mit zu den zentralen Themen quantitativ orientierter historischer Sozialforschung gehört. Von daher ergeben sich eine Fülle von Anknüpfungspunkten im Text.

Alle Auswertungen wurden mit dem Programmpaket SAS sowie mit dem Programm TDA erstellt. TDA ist ein von Götz Rohwer erstelltes Programm, das speziell auf die Probleme der Verlaufsdatenanalyse zugeschnitten ist. Es wird vom Autor zusammen mit einem ausgezeichneten Manual kostenlos vertrieben.

Nach sieben Jahren spiegelt die heute vorgelegte Fassung nicht nur den Wandel der Sichtweisen des Autors wider, sondern demonstriert auch die Veränderungen auf dem Sektor der Textverarbeitung. Dies ist im wesentlichen das Verdienst von Volker Verrel, der seinen guten alten EUMEL bis zum Letzten ausgereizt hat, ehe dieses Betriebs– und Programmsystem als Public–Domain Produkt auslaufen wird. Bei der sonstigen textlichen und graphischen Gestaltung haben mich Renate Bendel, Gisela Diekmeier und Hartmut Popken unterstützt. Alle verbliebenen Fehler sind – wie immer – Schuld des Autors und werden in der dritten Auflage bereinigt.

Bielefeld, im März 1992

# 1. Einführung

Da man bei Längsschnittdaten üblicherweise an Zeitreihen oder Paneldaten denkt, bedarf der Begriff "Verlaufsdaten" einer näheren Erläuterung. *Verlaufsdaten* sind ganz allgemein Informationen über die zeitliche Abfolge von Ereignissen. In Abschnitt 1.1 möchte ich an Hand von Beispielen erläutern, worum es sich dabei konkret handelt, welche neuen Erkenntnisse man mit Verlaufsdaten gewinnen kann und welche Fragestellungen im Rahmen einer Verlaufsanalyse untersucht werden können. In Abschnitt 1.2 werden Verlaufsdaten dann genauer definiert und von anderen Datenstrukturen (z.B. Zeitreihen) abgegrenzt. Abschnitt 1.3 beschäftigt sich schließlich mit Erhebungsproblemen.

Nach dieser Einführung konzentrieren sich dann alle folgenden Kapitel auf die eigentliche Auswertung. Kapitel 2 erläutert einige statistische Grundbegriffe, die für das Verständnis der Methode notwendig sind. In Kapitel 3 werden zwei Datensätze vorgestellt, an denen beispielhaft die folgende Auswertung demonstriert werden soll. Dabei werden auch die praktischen Probleme des Datenmanagements und der Auswertung mit Hilfe von EDV-Programmen angesprochen.[1] Die Auswertung selbst wird in mehr explorative und mehr konfirmatorische Verfahren unterteilt. In Kapitel 4 geht es darum, Einzelfragen auf einer mehr deskriptiven Ebene zu untersuchen, während in Kapitel 5 umfassende Erklärungsmodelle für die Daten getestet werden. Das Buch schließt mit einem Kapitel 6 über weiterführende Fragestellungen der Verlaufsdatenanalyse.

Alle wesentlichen Methoden sollen in einer Form behandelt werden, daß sie erstens an Hand eines einfachen Beispiels nachgerechnet werden können und daß zweitens ihre spezifische Anwendung in den Sozialwissenschaften deutlich wird. Für die praktische Auswertung stehen dann EDV-Programme zur Verfügung, jedoch sollte man nach der Lektüre dieses Textes in der Lage sein, einfache Probleme auch mit dem Taschenrechner zu lösen. Wer darüber hinaus an Spezialfragen oder den allgemeinen statistischen Eigenschaften der verwendeten Methoden interessiert ist, erhält auf diese Weise (hoffentlich) einen Einstieg in die statistische Fachliteratur.

---

1) Für die im Text verwendeten Beispiele sind die entsprechenden Programme in Anhang E dokumentiert.

Soziologische Anwendungen der Verlaufsdatenanalyse sind vor allem mit den Namen J. Coleman, M.T. Hannan und N.B. Tuma verbunden. Von ihnen stammen auch umfangreiche Monographien (COLEMAN 1981, TUMA/HANNAN 1984), die auf diesen Anwenderkreis zugeschnitten sind. In der Reihe "Quantitative Applications in the Social Sciences" informiert der Text von ALLISON (1984) über wesentliche Vorgehensweisen und ihre programmtechnische Umsetzung. Eine deutschsprachige Einführung in die Methoden der Verlaufsdatenanalyse lieferten zuerst DIEKMANN und MITTER (1984). Der ebenfalls deutschsprachige Standardtext von BLOSSFELD, HAMERLE und MAYER (1986) erläutert in umfassender Form statistische Theorie und anwendungsbezogene Datenanalyse. Beide Bücher demonstrieren auch die Umsetzung mit sozialwissenschaftlichen Programmpaketen. Die Monographien von KALBFLEISCH/PRENTICE (1980), LAWLESS (1982) und COX/OAKES (1984) sind die einschlägigen statistischen Fachbücher und für den sozialwissenschaftlichen Anwender daher nicht immer leicht zugänglich. In eine ähnliche Kategorie fällt die Monographie von LANCASTER (1990), der vor allem ökonometrische Anwendungen diskutiert. Weitere Lehrbücher stammen von CHIANG (1975), LEE (1980), ELANDT-JOHNSON/JOHNSON (1980), CROSS/CLARK (1975), LEE (1980), NELSON (1982), MANN et al. (1974), MILLER (1981) und HAMERLE/TUTZ (1989).

Dieser Text richtet sich an Personen, die bereits Grundkenntnisse statistischer (insbes. multivariater) Methoden haben. Eine gewisse Vertrautheit mit der einfachen linearen Regression ist Voraussetzung und erleichtert das Verständnis der hier zu besprechenden Methoden, da sie häufig in Analogie zu diesem klassischen Modell sozialwissenschaftlicher Datenanalyse entwikkelt werden. Vorausgesetzt werden auch Grundkenntnisse der Matrizenrechnung, die jedoch in der Regel nicht über die Multiplikation von Matrizen hinausgehen.

Methoden der Verlaufsdatenanalyse beruhen im wesentlichen auf Anwendungen der Wahrscheinlichkeitsrechnung (stochastische Prozesse). Begriffe wie Zufallsvariable, Erwartungswert, Dichte und Verteilungsfunktion sollten daher bekannt sein. Für den interessierten Leser findet sich dazu in Anhang A noch einmal eine Wiederholung der zentralen Begriffe. Außerdem werden die mathematischen Hilfsmittel erläutert, die in diesem Zusammenhang wichtig sind (Differential- und Integralrechnung). Ein Glossar in Anhang C erfaßt alle im Text verwendeten mathematischen Symbole.

## *1.1 Anwendungen der Verlaufsdatenanalyse*

Traditionell wurden die Methoden der Verlaufsdatenanalyse für medizinische (Survival Analysis) und technische Fragestellungen (Reliability Analysis) verwendet. Seit einiger Zeit kann man jedoch auch eine zunehmende Verbreitung in anderen Wissenschaftsbereichen konstatieren, so z.B. in der Soziologie, der Ökonomie, der Geschichtswissenschaft und der Politologie. Einige Anwendungsbeispiele sollen daher einen ersten Überblick über die substanzwissenschaftlichen Fragestellungen vermitteln, die mit Verlaufsdaten bearbeitet werden können.

### *1.1.1 Einige Beispiele*

Eine Standardanwendung der Survival Analysis findet sich z.B. in einer arbeitsmedizinischen Untersuchung von DYER (1975). Ziel dieser Längsschnittstudie war es, einige der bekanntesten Risikofaktoren wie z.B. Cholesterinspiegel des Blutes, Bluthochdruck und Zigarettenkonsum im Zusammenhang zu untersuchen. Die Stichprobe bestand aus 1233 Beschäftigten eines Chicagoer Energieunternehmens. Zu Beginn der Untersuchung waren die beteiligten Personen zwischen 40 und 59 Jahre alt und hatten keinerlei Herzkrankheiten. Während der Untersuchungsdauer von insgesamt 14 Jahren verstarben 246 Personen. Um den Einfluß der 3 Risikofaktoren gegeneinander abzuwägen, betrachtet DYER in einem multivariaten Modell die Wahrscheinlichkeit, einen bestimmten Zeitpunkt zu überleben. Grundlage seiner Berechnungen sind die Sterbedaten der untersuchten Personen. Dabei muß natürlich das Lebensalter berücksichtigt werden. Es zeigt sich, daß alle Faktoren außer dem Cholesterinspiegel signifikant das Sterberisiko erhöhen. Das Lebensalter rangiert dabei an erster Stelle gefolgt von Zigarettenkonsum und Bluthochdruck.

In der Soziologie sind diese Methoden vor allen Dingen durch einen Aufsatz zur dynamischen Analyse von "Event-Histories" bekannt geworden (TUMA/HANNAN/GROENEVELD 1979). Diese Arbeit ist im Zusammenhang mit einer Evaluierungsstudie entstanden, in der die Wirkung von negativen Einkommenssteuern als Instrument der Sozialhilfe untersucht werden sollte (Denver and Seattle Income Maintenance Experiment). Dabei erhalten Personen ohne eigenes Einkommen finanzielle Unterstützungsleistungen, die jedoch in dem Maße besteuert werden, in dem andere Finanzierungsquellen (z.B. eigenes Einkommen, andere Sozialhilfen)

hinzukommen. Die Besonderheit dieses Hilfsprogramms besteht also darin, daß unterstützungsberechtigte Personen durchaus eigenes Einkommen beziehen können, Beihilfen jedoch dementsprechend gemindert (besteuert) werden. Durch die Kombination von Transferzahlungen und Eigenleistungen wird also quasi ein Mindesteinkommen gesichert. Dagegen sind bei den üblichen US-amerikanischen Sozialhilfeprogrammen (z.B. food stamps, aid for dependent children) Leistungen daran geknüpft, daß die Bezieher dieser Leistungen kein eigenes Einkommen haben. Ein unerwünschter Nebeneffekt dieser Regelung ist jedoch, daß Einkommensbezieher (in der Regel der Mann) zumindest formell den Haushalt verlassen, damit die restlichen Haushaltsmitglieder bezugsberechtigt werden. In den entsprechenden Risikopopulationen war daher eine Zunahme der Familien mit weiblichem Haushaltsvorstand zu beobachten, so daß sich indirekt die Kosten des Wohlfahrtssystems erhöhten.

Auf dem Hintergrund dieser Überlegungen erklärt sich das Interesse der Autoren, die Stabilität von Ehen zu untersuchen (vgl. HANNAN/TUMA/GROENEVELD 1977; dort wird auch das Design des Feldexperiments näher beschrieben). Sie argumentieren, daß die Kosten öffentlicher Transferzahlungen nur dann angemessen beurteilt werden können, wenn u.a. die Wirkungen auf das Heiratsverhalten bekannt sind. Leider gibt es ihrer Ansicht nach zu dieser Frage nur wenig gesichertes Wissen. Die meisten Forschungen zur Heiratsmobilität würden auf Querschnittsuntersuchungen beruhen und den Einfluß kultureller und sozialstruktureller Merkmale (soziale Nähe/Distanz der Partner) betonen. TUMA, HANNAN und GROENEVELD wählen daher einen explizit dynamischen Ansatz, um den Einfluß situationsabhängiger und damit potentiell veränderlicher Merkmale untersuchen zu können. In diesem Rahmen können sie dann den Effekt von Einkommensschwankungen testen, die sich im Zusammenhang mit o.g. öffentlichen Sozialleistungen ergeben.

Konkret untersuchen sie die Eheschließungs- bzw. Scheidungsraten der an dem Experiment beteiligten Frauen. Dazu verwenden sie Angaben über die Ehedauer (bei Verheirateten) bzw. den Zeitpunkt der Eheschließung (bei Ledigen oder Geschiedenen). Darüber hinaus werden zusätzliche Einflußfaktoren kontrolliert, darunter Einkommen, Anzahl und Alter der Kinder sowie Ausbildung und Alter der Frau. Es zeigt sich, daß Unterstützungszahlungen in Form negativer Einkommenssteuern ebenfalls die Instabilität von Ehen erhöhen, wobei dieser Effekt jedoch im Zeitablauf variiert.

Genauso wie der Familienstatus verändern sich viele andere soziale Tatbestände im Zeitablauf. Während das in der sozialwissenschaftlichen Theoriebildung schon lange berücksichtigt wird (vgl. dynamische Begriffe wie "Mobilität", "sozialer Wandel", "Evolution" etc.), beruhen empirische Forschungen noch weitgehend auf statischen Analyseverfahren. Dabei könnte eine Berücksichtigung der Zeitdimension viele zusätzliche Informationen liefern.

Ich habe versucht, diesen zusätzlichen Informationsgewinn an einem Beispiel aus der Kriminologie zu zeigen, das sich mit Determinanten der Rückfälligkeit von Strafentlassenen beschäftigte (1984a). Grundlage meiner Auswertung waren die Aufzeichnungen eines Sozialarbeiters über die von ihm betreuten Personen. Diese Daten geben Auskunft über finanzielle Belastungen der Mandanten sowie über Probleme bei der Arbeits- und Wohnungssuche und im sozialen Umfeld (Verwandte, Bekannte, Freunde). Erwartungsgemäß hängt die erfolgreiche Bewährung der Person von den genannten Problemfeldern ab. Berücksichtigt man jedoch zusätzlich den Zeitpunkt einer eventuellen Rückfälligkeit, dann zeigt sich, daß die untersuchten Personen zu Beginn der Bewährungszeit besonders gefährdet sind. Es liegt daher nahe, in den ersten Wochen nach der Haftentlassung eine besondere Betreuung vorzusehen.

In diesem Sinne kann man für viele andere Wissensbereiche zeigen, daß mit Verlaufsanalysen fruchtbare Forschungsergebnisse zu erzielen sind:

- Arbeitslosigkeit und Beschäftigung im Lebenszyklus (DIPRETE 1981).
- Einfluß persönlicher Ressourcen und materieller Gratifikationen auf freiwillige Beschäftigungswechsel (TUMA 1976).
- Klassische und segmentationstheoretische Erklärungsansätze der Arbeitsmobilität (TUMA 1982a, SÖRENSEN/TUMA 1978).
- Einfluß wirtschaftlicher Rahmenbedingungen und persönlicher Merkmale auf soziale Auf- und Abstiege (ANDRESS 1983).
- Politische Stabilität verschiedener Nationalstaaten (HANNAN/ CARROLL 1981).
- Gesundheitsbelastungen und Arbeitsmobilität (ANDRESS 1980).
- Stabilität von Firmen und Organisationen in der amerikanischen Zeitungsindustrie (FREEMAN et al. 1983).
- Einfluß von Organisationsmerkmalen auf die Karrieremobilität (CARROLL/ MAYER 1984).

- Stabilität von Armutsrisiken am Beispiel von AFDC–Bezieherinnen (PLOTNICK 1983).
- Determinanten der Arbeitslosigkeitsdauer (SCHNEIDER 1988).
- Job–Search–Theorie und Stellensuchdauer von Hochschulabsolventen (ZIEGLER et al. 1988).
- Arbeitsplatzwechsel und Beschäftigungsstabilität in einem Großunternehmen (DIEKMANN/PREISENDÖRFER 1988).
- Einfluß der schulischen Bildung und der Bildungsexpansion auf das Heiratsverhalten (DIEKMANN 1990).
- Arbeitsmärkte im öffentlichen Dienst und in der Privatwirtschaft aus der Perspektive von Berufsverläufen (BECKER 1990).
- Kohortendifferenzierung und Karriereprozeß (BLOSSFELD 1989).
- Determinanten der Mandatsdauer von Abgeordneten der deutschen Reichstage 1867–1918 (ANDRESS et al. 1992).

Diese relativ willkürliche Liste von Beispielen kann noch beliebig weitergeführt werden, doch zeigt schon dieser erste Überblick die Vielfalt der Anwendungsmöglichkeiten.

Allen ist gemeinsam, daß Tatbestände untersucht werden, die sich im Zeitablauf ändern (Gesundheitszustand, Familienstatus, Erwerbstätigkeit, Beruf, Bewährung usw.). Man spricht auch von *sozialen Prozessen.* Die verwendeten Daten bestehen in der Regel aus Angaben über Veränderungszeitpunkte der untersuchten Merkmale (Heiratsdatum, Ehedauer, Arbeitslosigkeitsdauer, Zeitpunkt einer Straftat oder eines Tätigkeitswechsels etc.). Veränderungen bezeichnet man auch als *Ereignisse* und spricht daher von Ereignisdaten bzw. Ereignisdatenanalyse. Eine Kette von Ereignissen konstituiert einen *Verlauf* – beispielsweise einen Berufsverlauf, der sich aus mehreren Berufswechseln zusammensetzt. Ich bevorzuge daher statt der Bezeichnung *Ereignisdatenanalyse* den umfassenderen Begriff *Verlaufsdatenanalyse.*

### *1.1.2 Warum Verlaufsdatenanalyse?*

Es steht natürlich außer Frage, daß Längsschnittuntersuchungen sehr viel mehr Informationen enthalten und damit auch sehr viel differenziertere Auswertungen ermöglichen. Auch besteht kein Zweifel daran, daß sich die meisten Merkmale im Zeitablauf verändern, so daß eine dynamische Analy-

se schon allein aus theoretischen Gründen einer statischen Analyse vorzuziehen ist. Angesichts der Kosten und des Mehraufwandes von Längsschnittuntersuchungen stellt sich jedoch die Frage, ob man nicht mit einfacheren Methoden zu ähnlichen Ergebnissen kommt, die zwar nicht so exakt und theoretisch befriedigend sind, die aber im wesentlichen die gleichen Schlußfolgerungen erlauben.

Genügt es also beispielsweise, am Ende des o.g. Feldexperiments zu fragen, wie viele Frauen ihren Familienstatus geändert haben, oder einfach nur zu untersuchen, wie viele Probanden während der Bewährungszeit rückfällig wurden, ohne den genauen Zeitpunkt des jeweiligen Ereignisses zu kennen? Möglicherweise ergibt sich ja schon bei dieser reduzierten Fragestellung, daß Frauen mit Unterstützungszahlungen instabilere Ehen führen und Personen mit sozialen Problemen häufiger rückfällig werden. Welchen zusätzlichen Informationsgewinn erzielt man also durch Verlaufsdaten und rechtfertigt dieser Vorteil den zusätzlichen Mittelaufwand?

In beiden Fällen läßt sich diese Frage bejahen, denn das untersuchte Merkmal (Familienstatus, Bewährung) unterliegt starken Schwankungen im Zeitablauf. Beidesmal zeigt sich, daß Ereignisse (Scheidungen, Heiraten, Straftaten) kurz nach Beginn der Untersuchung sehr viel häufiger auftreten als zu späteren Zeitpunkten. Anders ausgedrückt, die Rückfallquote ist kurz nach Haftentlassung sehr viel höher. Das gleiche gilt für die Stabilität von Ehen, die offenbar bei Eintritt in das Feldexperiment sehr viel stärker gefährdet werden als in späteren Stadien der Untersuchung.

Diese Beobachtung ist ein bekanntes Phänomen sozialer Prozesse, die durch einen äußeren Eingriff in ihren bisherigen Bewegungsgesetzen beeinflußt werden und erst nach einiger Zeit in ihren Ausgangszustand zurückkehren. Bei Querschnittsanalysen wird ein solcher Gleichgewichtszustand unterstellt. Man sollte aber die zeitliche Veränderung berücksichtigen, denn daraus ergeben sich unmittelbare Konsequenzen für das praktische Handeln. Z.B. wird man bei den ehemaligen Häftlingen in den ersten Wochen nach Haftentlassung eine besondere Betreuung vorsehen – und zwar unabhängig von den sonstigen Problemen der Häftlinge.

Die Beeinflussung eines sozialen Prozesses durch externe Eingriffe deutet darauf hin, daß die Ursache eines im Zeitablauf variierenden Scheidungs– oder Rückfallrisikos darin zu suchen ist, daß sich bestimmte Bedingungen verändern (Haushaltseinkommen bei Transferzahlungen, Lebensumstände nach Haftentlassung), die das untersuchte Merkmal beeinflussen. Anders formuliert, die erklärenden Variablen sind selbst zeitabhängig und

gerade diese Veränderungen in den Ursachenkonstellationen können eben nicht mit Querschnittsdaten abgebildet werden. Ein klassisches Beispiel, das in diesem Zusammenhang häufig zitiert wird und in fast allen Statistik-Lehrbüchern zur Verlaufsdatenanalyse zu finden ist, ist die Stanforder Herz-Transplantations-Studie (CROWLEY/HU 1977). Dabei ging es vor allem um die Effekte einer möglichen Herz-Transplantation auf die Überlebenszeit der Patienten. Die Besonderheit dieser Studie besteht darin, daß ein Teil der Untersuchungspersonen im Verlauf der Untersuchung seinen Status wechselt: Ein Patient rechnet so lange zur Kontrollgruppe, bis ein geeigneter Spender gefunden wird und die Transplantation durchgeführt werden kann. Von diesem Zeitpunkt an gehört er der Experimentalgruppe an. Wenn die Transplantation einen lebensverlängernden Effekt haben soll, dann muß die Überlebenszeit der Experimentalgruppe signifikant länger sein. Jede einigermaßen aussagekräftige Auswertungsstrategie muß zur Beantwortung dieser Frage neben anderen Risikofaktoren den Zeitpunkt der Transplantation berücksichtigen und dieses ist, wie wir sehen werden, im Rahmen einer Verlaufsanalyse sehr viel präziser und einfacher möglich, als im Rahmen irgendeines anderen Untersuchungsdesigns.

Wenn sich die untersuchten (abhängigen und unabhängigen) Merkmale im Zeitablauf ändern, hat das natürlich auch Auswirkungen auf die Erhebung der Daten. Beschränkt man sich auf Querschnittsdaten, dann hängen die Ergebnisse sehr stark davon ab, zu welchem Zeitpunkt die Erhebung durchgeführt wurde. Auch durch einfache Erweiterungen des Querschnittsdesigns (z.B. vorher/nachher-Messungen) können diese Defizite nur bedingt aufgehoben werden. SINGER und SPILERMAN (1974, 1976a,b) zeigen, daß durch Daten über zwei Zeitpunkte nicht genügend Informationen zur Verfügung stehen, um eindeutig einen bestimmten Veränderungsprozeß identifizieren zu können. Ähnlich wie man zwei Punkte in einem Koordinatenkreuz durch beliebige Kurven verbinden kann, können zwischen zwei Zeitpunkten die unterschiedlichsten Veränderungen stattfinden.

Kenntnis der zeitlichen Veränderung verbessert auch die Prognosefähigkeit der verwendeten Modelle. Das kann man sich relativ leicht an dem arbeitsmedizinischen Beispiel deutlich machen. Hier ergab ein statisches Modell, das lediglich die Verhältnisse am Ende der Untersuchung (also nach $t=14$ Jahren) berücksichtigte, strukturell ähnliche Ergebnisse wie das dynamische Modell, das die Veränderung der Mortalität im Verlauf der Untersuchung (d.h. mit fortschreitendem Alter der Untersuchungspersonen) betrachtete (vgl. DYER 1975). Die letzte Information benötigt man jedoch, wenn man Aussagen über die Mortalität nach weiteren $\Delta t$ Untersuchungs-

jahren machen will. Dann genügt es nicht, den Effekt des Alters, wie er sich z.B. in einem klassischen Regressionsmodell ergibt, linear fortzuschreiben. Ein nicht-linearer Verlauf ist wahrscheinlicher, kann aber nur dann eindeutig diagnostiziert werden, wenn man die im Untersuchungsablauf beobachteten Veränderungen berücksichtigt.

Dieses Plädoyer für dynamische Analysen darf jedoch nicht in der Weise mißverstanden werden, daß in allen Fällen möglichst viele zeitbezogene Informationen erhoben und ausgewertet werden sollen. Es versteht sich eher als Gegenrede gegen Positionen, die zeitbezogene Fragestellungen weitgehend ignorieren wollen. Jede Datenanalyse beginnt natürlich mit einfachen, meist querschnittsbezogenen Auswertungen, sollte aber, soweit es möglich und theoretisch sinnvoll ist, im weiteren Verlauf der Untersuchung alle zur Verfügung stehenden zeitbezogenen Informationen berücksichtigen. In vielen Fällen ergeben sich dadurch zusätzliche Erkenntnisse, wenn auch der relative Gewinn dieser Strategie von Fall zu Fall erneut begründet werden muß. Außerdem ist der praktische Aufwand zur Durchführung einer Verlaufsanalyse mit den heute zur Verfügung stehenden EDV-Programmen nicht wesentlich höher als bei üblichen Kreuztabellen- oder Korrelationsanalysen. Welche Untersuchungsgegenstände (abhängiges Merkmal) und Hypothesen (unabhängige Merkmale) in einer solchen Untersuchung bearbeitet werden können, soll an Hand des folgenden, etwas konkreter ausgeführten Beispiels deutlich werden.

### *1.1.3 Untersuchungsgegenstände*

Jeder hat sicherlich schon einmal einen Lebenslauf für eine Bewerbung geschrieben und daher soll uns dieses alltagsweltliche Beispiel zur Illustration typischer Untersuchungsfragen im Rahmen einer Verlaufsanalyse dienen. Tabelle 1.1 zeigt den Berufsverlauf eines gewissen Herrn Müller. Dieser Lebenslauf erscheint relativ genau und ich will auch einmal annehmen, daß diese Daten eine exakte Rekonstruktion des beruflichen Lebenslaufes von Herrn Müller ermöglichen. Daß Verlaufsdaten sehr viel ungenauer erhoben werden können und auch dieses Beispiel möglicherweise ein sehr ungenaues Abbild der tatsächlichen Veränderungen darstellt, ist mir bewußt, doch will ich dieses Problem erst später diskutieren. Zunächst gehe ich einmal davon aus, daß mir alle notwendigen Informationen zur Verfügung stehen.

Wie man weiter sieht, sind Zahl und Umfang der erhobenen Informationen relativ groß, obwohl in der Tabelle eigentlich nur Daten über das abhängige Merkmal enthalten sind. Würde man darüber hinaus die unabhängigen Merkmale und deren Veränderung berücksichtigen, dann würde der Datensatz von Herrn Müller noch um einiges größer.

*Tabelle 1.1: Tabellarischer Lebenslauf von Herrn Müller*

| Kalender– jahr | Jahre nach Berufseintritt | Tätigkeit | Wirtschaftszweig |
|---|---|---|---|
| vor 1961 | – | Berufsausbildung | verarbeitendes Gewerbe |
| 1961 | 0 | angelernter Arbeiter | verarbeitendes Gewerbe |
| 1963 | 2 | Facharbeiter | verarbeitendes Gewerbe |
| 1965 | 4 | angelernter Arbeiter | Baugewerbe |
| 1967 | 6 | Facharbeiter | Baugewerbe |
| 1969 | 8 | Meister | Baugewerbe |
| 1970 | 9 | einf. Angestellter | Staat |

Bei einer empirischen Untersuchung betrachtet man nun mehrere solcher Berufsverläufe. Angesichts des Umfangs der erhobenen Information ist es erst einmal sinnvoll, das *Datenmaterial zu selektieren und zu aggregieren*:

a) Zunächst könnte man sich mit der Frage beschäftigen, wann eine Person das erste Mal ihren Beruf wechselt. Dazu würde man lediglich die jeweils *ersten Tätigkeiten* betrachten, ohne zu berücksichtigen, wie die jeweilige Tätigkeit beendet wird.

b) Daran anschließend kann man untersuchen, welche *Art von Wechsel* stattfindet. Da die einzelnen Tätigkeiten offenbar sehr differenziert erhoben wurden, sind alle möglichen Wechsel zwischen Tätigkeiten und Wirtschaftszweigen denkbar. Um zu kleine Fallzahlen bei einzelnen Übergängen zu vermeiden, wird es daher notwendig sein, verschiedene Arten von Wechseln zusammenzufassen. Beispiele wären etwa: berufliche Auf– versus berufliche Abstiege; Wechsel mit ausschließlicher Veränderung der Tätigkeit versus Wechsel mit gleichzeitiger Veränderung von Tätigkeit und Wirtschaftszweig usw.

c) Schließlich kann man *alle Tätigkeiten* gemeinsam betrachten, um die Tatsache zu berücksichtigen, daß sich Tätigkeitswechsel im Laufe eines Berufslebens wiederholen können. Damit werden alle Informationen der erhobenen Berufsverläufe verwertet. Auch hier wird es nötig sein, verschiedene Arten von Wechseln zusammenzufassen.

Je nachdem welche Informationen des Urmaterials für die folgende Auswertung berücksichtigt werden sollen, wird man also Daten unterschiedlichen Umfangs generieren: Beispielweise eine Datei A, die nur die ersten Tätigkeiten enthält. Oder eine Datei B, die nur die ersten Tätigkeiten erfaßt, diese aber nach der Art des Wechsels differenziert (z.B. Abstiege vs. Aufstiege). Schließlich eine Datei C, in der alle Tätigkeiten mit ihrem jeweiligen Abschluß festgehalten sind.

Auswahl und Aggregation der Daten hängen natürlich eng mit der gewählten Fragestellung der Untersuchung zusammen. Angesichts der Reichhaltigkeit des Urmaterials sind die verschiedensten Untersuchungsfragen denkbar. Ich möchte nur einige andeuten und die entsprechenden Datenquellen nennen:

- Wie lange dauert im Durchschnitt die erste Tätigkeit (Datei A)?
- Wie groß ist die Wahrscheinlichkeit, 5 Jahre nach Berufseintritt noch immer im Erstberuf zu arbeiten (Datei A)?
- Wenn ein Wechsel stattfindet, in welchem Jahr ist er am wahrscheinlichsten (Datei A)?
- Welche Tätigkeit folgt am wahrscheinlichsten auf den Erstberuf (Datei B)?
- Hängt die Art des Wechsels vom Erstberuf ab (Datei B)?
- Wie lange dauern alle Tätigkeiten im Durchschnitt und wann sind Wechsel am wahrscheinlichsten (Datei C)?
- Wieviel Tätigkeitswechsel ereignen sich durchschnittlich in den ersten 10 Berufsjahren (Datei C)?
- Wenn man von einer Kohorte von Berufsanfängern und den von ihnen gewählten Erstberufen ausgeht, wie verteilt sich diese Personengruppe 10 Jahre nach Berufseintritt auf die verschiedenen Tätigkeiten (Datei C)?

*Abhängige Merkmale einer Verlaufsdatenanalyse* können also sein: Zustandsdauer, Wahrscheinlichkeit von Ereignissen, Zeitraum ohne Ereignis, Übergangswahrscheinlichkeiten zwischen verschiedenen Zuständen, Anzahl

Ereignisse pro Zeitintervall, Zustandsverteilung zu einem gegebenen Zeitpunkt. Wie können diese Zielvariablen nun erklärt werden? Im folgenden Abschnitt möchte ich dazu einige prototypische Hypothesen vorstellen.

### *1.1.4 Hypothesen*

Egal welches abhängige Merkmal man betrachtet, bei der *Erklärung sozialer Prozesse* lassen sich grob fünf Hypothesengruppen unterscheiden:

1. zustandsabhängige Prozesse,
2. heterogene Prozesse,
3. zeitabhängige Prozesse,
4. Einflüsse der Vorgeschichte,
5. unberücksichtigte Einflüsse und Meßfehler.

Alle in den obigen Anwendungsbeispielen untersuchten Einflußfaktoren lassen sich diesen Oberbegriffen zuordnen.

Betrachten wir ganz allgemein berufliche Mobilitätsprozesse, die sich formal als Wechsel von einem (Ausgangs-)Zustand in einen neuen (Ziel-)Zustand charakterisieren lassen, dann sind Wechsel zwischen den verschiedenen Zuständen in der Regel nicht gleich wahrscheinlich und man spricht von einem *zustandsabhängigen Prozeß*. Klassifiziert man beispielsweise Tätigkeitswechsel mit Hilfe einer Statusskala in Aufstiege, Abstiege und horizontale Mobilität, dann wird man erwarten, daß die Zustandswechsel, die als Abstieg eingeordnet wurden, sehr viel weniger wahrscheinlich sind als Aufstiege und horizontale Mobilität.

Darüber hinaus ist sicherlich davon auszugehen, daß das Abstiegsrisiko je nach Merkmalen der Person (z.B. Qualifikation) oder der Wirtschaftszweige (z.B. Personalbedarf) variiert. Die Untersuchungsgruppe ist also im Hinblick auf das abhängige Merkmal heterogen, d.h. die Wahrscheinlichkeit bestimmter Ereignisse variiert mit den Merkmalen der Untersuchungspersonen (*heterogener Prozeß*).

Schließlich ist anzunehmen, daß die Wahrscheinlichkeit eines Abstiegs mit der Dauer der Tätigkeit abnimmt, da die Person spezifische Qualifikationen oder Sozialansprüche erwirbt, die dieses Risiko verringern. Die Wahrscheinlichkeit eines bestimmten Ereignisses verändert sich also im Zeitablauf (*zeitabhängiger Prozeß*).

Um diese drei Hypothesengruppen überprüfen zu können, muß man nicht notwendigerweise alle Einzelheiten des Urmaterials berücksichtigen. Hierzu genügt es, sich auf die jeweils ersten Tätigkeiten zu beschränken (Datei A oder B), vorausgesetzt man kann davon ausgehen, daß die ersten Tätigkeiten ein repräsentatives Abbild des gesamten Berufsverlaufs darstellen. Will man hingegen Einflüsse der Vorgeschichte untersuchen, dann muß man mehrere Tätigkeitswechsel, also idealiter den gesamten Berufsverlauf einer Person berücksichtigen (Datei C).

Man könnte z.B. argumentieren, daß die Wahrscheinlichkeit eines Abstiegs mit der Anzahl vorheriger Abstiege zunimmt, weil entweder die Qualifikation der Person sinkt oder eine gewisse Stigmatisierung stattfindet. Einen ähnlichen negativen Effekt dürfte auch die Beschäftigungsdauer in früheren niedrig qualifizierten Positionen haben. Umgekehrt ließe sich einwenden, je länger eine Person überhaupt erwerbstätig ist, um so größer sind ihre allgemeinen Berufserfahrungen und um so geringer daher ihr Abstiegsrisiko. In allen drei Fällen wird jedenfalls ein Einfluß der *Vorgeschichte* auf das Auftreten weiterer Ereignisse vermutet, wobei im ersten Beispiel die *Häufigkeit früherer Ereignisse* und in den beiden anderen Beispielen die (summierte) *Dauer früherer Zustände* eine Rolle spielen.

Ein vollständiges Erklärungsmodell wird natürlich Kombinationen dieser 5 Hypothesengruppen verwenden. Man könnte z.B. annehmen, daß sich das Abstiegsrisiko für ältere Arbeitnehmer nicht wesentlich verändert. Also wird man den oben beschriebenen zeitabhängigen Prozeß nur für jüngere Arbeitnehmer unterstellen, für die älteren dagegen von einem zeitkonstanten Abstiegsrisiko ausgehen. Folglich wird man einen Interaktionseffekt zwischen Heterogenitätsfaktoren (hier: Alter) und Zeitabhängigkeit in das Modell aufnehmen.

Auf diese Weise kann man beliebig komplizierte Modelle konstruieren, doch reichen unsere Theorien in der Regel nicht aus, um die Daten vollständig zu erklären – ganz abgesehen davon, daß die Daten möglicherweise fehlerhaft sind. Wie in üblichen Regressionsmodellen sollten daher Erklärungsmodelle für soziale Prozesse einen *Fehlerterm* enthalten, der unberücksichtigte Einflüsse *(unbekannte Heterogenität* mangels Theorien oder Daten) sowie *Meßfehler* erfaßt. Jedoch sei schon an dieser Stelle darauf hingewiesen, daß die Verwendung eines Fehlerterms in Erklärungsmodellen für Verlaufsdaten sehr viel mehr Schwierigkeiten bereitet als in üblichen Regressionsmodellen.

## *1.2 Modellierung zeitkontinuierlicher Veränderungsprozesse mit Verlaufsdaten*

Bei Verlaufsdaten handelt es sich um eine spezifische Form von Längsschnittdaten, die in diesem Abschnitt von anderen Typen zeitbezogener Informationen wie etwa Zeitreihen oder Paneldaten abgegrenzt werden soll. Verlaufsdaten gehen von einer zeitkontinuierlichen Erhebung aus, während sich die anderen Längsschnittdesigns u.a. dadurch auszeichnen, daß Veränderungen nur zu diskreten Zeitpunkten erfaßt werden. Dabei stellt sich natürlich die Frage, ob Veränderungsprozesse, wie sie in den Sozialwissenschaften üblicherweise untersucht werden, überhaupt einem bestimmten Zeitschema unterliegen, so daß ein zeitdiskretes Erhebungsdesign gerechtfertigt wäre.

### *1.2.1 Längsschnittdaten*

Sozialwissenschaftliche Daten können drei verschiedene Dimensionen erfassen: Untersuchungseinheiten, Variablen und Zeitpunkte. Die resultierende Datenmatrix ist in Abbildung 1.1 dargestellt (vgl. CATTELL 1946, 1952; BUSS 1974). Querschnitts– oder Zeitreihendaten sind als Subdesigns in dieser Datenmatrix enthalten:

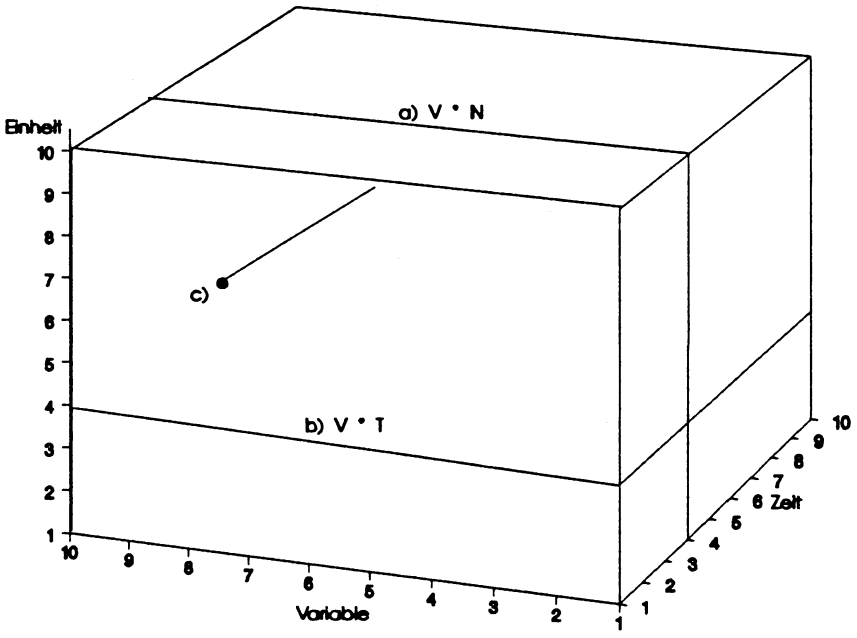
a) Wenn man verschiedene Merkmale *zu einem Zeitpunkt* bei verschiedenen Untersuchungseinheiten erhebt, erhält man die zweidimensionale Datenmatrix V·N, die durch den senkrechten Schnitt a durch den Datenwürfel dargestellt ist (*Querschnittsdaten*).

b) Wenn man *bei einer Untersuchungseinheit* mehrere Merkmale im Zeitablauf erfaßt, entsteht umgekehrt die zweidimensionale Datenmatrix V·T, die durch den waagerechten Schnitt b durch den Datenwürfel dargestellt ist (*Zeitreihendaten*).

Querschnittsdaten sind überall dort angebracht, wo praktisch keine Veränderungen im Zeitablauf auftreten oder wo die Unterschiede im Zeitablauf für nicht so bedeutend gehalten werden, wie die zwischen den Untersuchungseinheiten. Für Zeitreihen gelten die umgekehrten Überlegungen. Hier konzentriert man sich auf die zeitliche Variation und vernachlässigt Unterschiede zwischen den Untersuchungseinheiten. In den Fällen aber, in denen sowohl Unterschiede im Zeitablauf als auch zwischen Unter-

suchungseinheiten eine Rolle spielen, bedarf es eines dreidimensionalen Forschungsdesigns, das Untersuchungseinheiten, Variablen und Zeitpunkte gleichermaßen erfaßt (vgl. den gesamten Datenwürfel).

*Abbildung 1.1: Datenwürfel*




Ich verwende den Begriff *Längsschnitt* – oder *zeitbezogene Daten* für alle die Forschungsdesigns, die mindestens die Zeitdimension des Datenwürfels erfassen: Also auch die Erhebung eines Merkmals bei einer Untersuchungseinheit (vgl. die Linie c des Datenwürfels). Je nachdem, wie viele Informationen zusätzlich zur Verfügung stehen, sind Datenanalysen zunehmenden Differenzierungsgrades möglich:

– Die Kenntnis des zeitlichen Verlaufs eines Merkmals (vgl. Linie c) ermöglicht nur die Beschreibung dieses konkreten Veränderungsprozesses.

- Liegen Informationen über andere Variablen und deren zeitliche Variation vor (vgl. Schnitt b), kann man diese Daten benutzen, um den Verlauf des ersten Merkmals zu erklären. Veränderung in einem Bereich ist damit ein Resultat von Veränderungen in anderen Bereichen.
- Bei Berücksichtigung verschiedener Untersuchungseinheiten (vgl. den gesamten Datenwürfel) kann man schließlich deren individuelle Merkmale und zeitliche Variation zur Erklärung heterogener Veränderungsprozesse heranziehen.

Alle Längsschnittanalysen sind jedoch nur dann sinnvoll, wenn das untersuchte Merkmal tatsächlich relevanten Veränderungen im Zeitablauf unterliegt. Sollte dieser zeitliche Wandel nur geringfügig sein, dann erübrigt sich die Erhebung zeitbezogener Daten und man kann auf gängige Querschnittsdaten zurückgreifen (vgl. Schnitt a).

Geht man z.B. davon aus, daß sich Werte und Normen einer Person nur sehr langfristig verändern, aber zwischen den Personen sehr große Unterschiede bestehen, dann ist es vollkommen ausreichend, eine Befragung von Personen durchzuführen, in der alle wesentlichen Merkmale zu einem Zeitpunkt erhoben werden. Ist man jedoch umgekehrt der Meinung, daß sich Werte und Normen sehr wohl im Zeitablauf verändern, dann wird man nicht umhin können, diese Veränderungen durch geeignete Erhebungsverfahren zu erfassen. Natürlich sind auch Problemstellungen denkbar, bei denen Unterschiede zwischen den Untersuchungseinheiten keine Rolle spielen. Das ist allerdings eher bei Makro– als bei Mikrophänomenen der Fall, wie etwa bei BRENNER's (1979) Untersuchungen über den Zusammenhang von Wirtschaftskrisen, Arbeitslosigkeit und psychischen Erkrankungen.

### *1.2.2 Zeitkontinuierliche und zeitdiskrete Veränderungsprozesse*

Längsschnittdaten erfassen also mindestens die Zeitdimension des Datenwürfels. Um die Spezifika von Verlaufsdaten zu verstehen, ist es notwendig, sich mit dieser Zeitachse etwas eingehender zu beschäftigen. Bei der Modellierung von Veränderungsprozessen stellt sich zunächst die Frage, *wann* die untersuchten Änderungen tatsächlich eintreten, wie also die untersuchte Zeitachse unter theoretischen Gesichtspunkten formalisiert werden kann. Statt einer Wiederholung der umfänglichen Debatte über zeitdiskrete und zeitkontinuierliche Modelle (vgl. vor allem die Arbeiten von J.

COLEMAN) möchte ich das Problem an Hand einiger Beispiele von Veränderungsprozessen illustrieren.

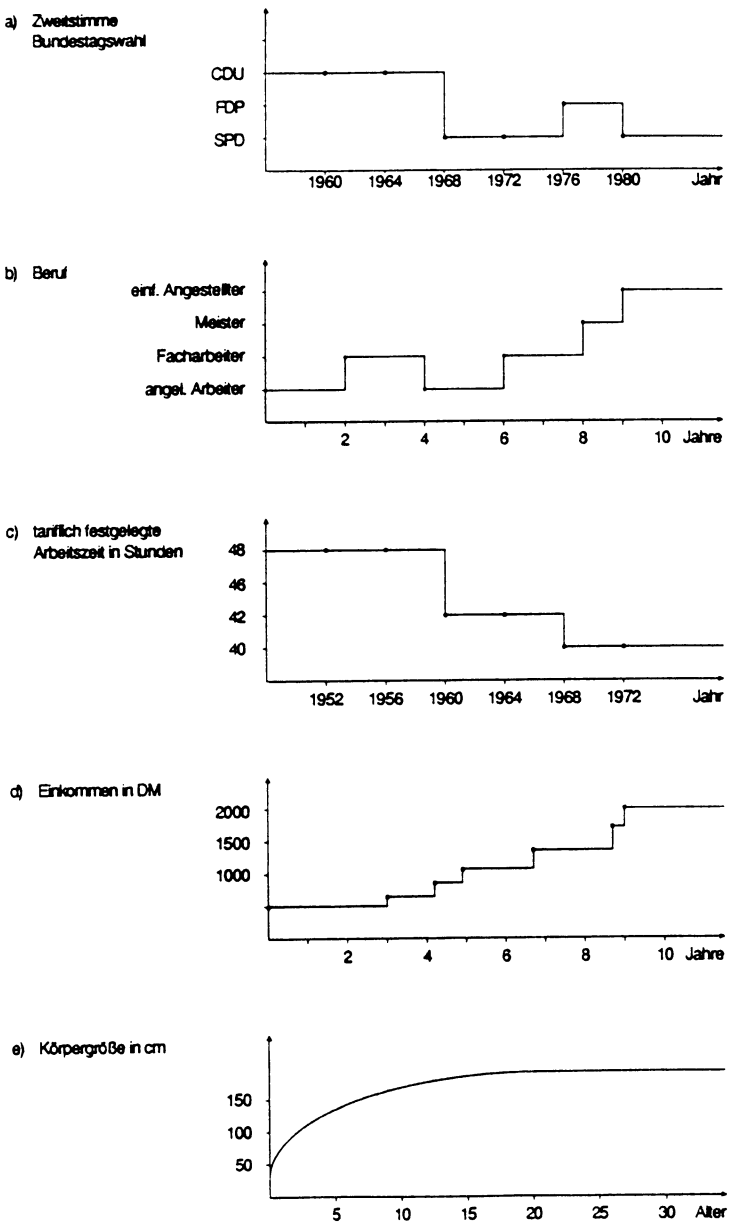
Betrachten wir dazu den einfachsten Fall einer Längsschnittanalyse: die Veränderungen eines Merkmals bei einer Untersuchungseinheit im Zeitablauf (vgl. Linie c). Erhebungsprobleme sollen uns zunächst nicht beschäftigen. Zu jedem Zeitpunkt des Untersuchungszeitraumes sei daher der Zustand der Einheit bekannt. Die Zahl der möglichen Zustände ist von den Ausprägungen des betrachteten Merkmals abhängig. Die einzelnen Zustände schließen einander aus, d.h. keine Untersuchungseinheit kann sich gleichzeitig in zwei verschiedenen Zuständen befinden. Alle auftretenden Zustände lassen sich eindeutig durch Zahlen identifizieren und der resultierende Set von Zahlen wird üblicherweise *Zustandsraum* genannt.

Im Zeitablauf durchwandert die Untersuchungseinheit mehrere Zustände. z(t) sei die Zahl, die den Zustand der Einheit zum Zeitpunkt t repräsentiert. Die geordnete Sequenz $z(t_0)$, $z(t_1)$, $z(t_2)$, ... beschreibt den Weg der Untersuchungseinheit durch den Zustandsraum. Man spricht auch von dem *zeitlichen Verlauf* der Zufallsvariablen Z(t). Dieser läßt sich in einem zweidimensionalen *Verlaufsdiagramm* darstellen, auf dessen Achsen die Zeit und die einzelnen Zustände abgetragen werden. Abbildung 1.2 zeigt fünf Beispiele solcher zeitlichen Verläufe mit Variablen unterschiedlichen Meßniveaus und unterschiedlichen Formen der Veränderung, darunter auch das obige Beispiel eines Berufsverlaufs (vgl. Tabelle 1.1).

Untersuchungseinheiten sind Personen und es geht um Veränderungen des Wahlverhaltens, der beruflichen Tätigkeit, der Arbeitszeit, des Einkommens sowie der Körpergröße. Die Beispiele zeigen sowohl Verläufe, bei denen Veränderungen *jederzeit* auftreten können, als auch Verläufe, bei denen Änderungen nur zu bestimmten *diskreten* Zeitpunkten auftreten können. Wenn man selber einmal versucht, diese Liste von Beispielen zu ergänzen, dann wird einem sehr schnell deutlich, daß es sehr schwierig ist, Beispiele des zweiten Typs zu finden. Unter theoretischen Gesichtspunkten scheint daher in vielen Fällen die Annahme eines zeitkontinuierlichen Prozesses die angemessenere Ausgangshypothese zu sein. Das gilt ganz besonders dann, wenn man Veränderungen auf der begrifflichen Ebene betrachtet.

Betrachten wir z.B. das theoretische Konstrukt "Parteipräferenz": Es ist anzunehmen, daß die Bevorzugung einer bestimmten Partei einem beständigen Wandel unterliegt. Die latente Variable "Parteipräferenz" kann sich also jederzeit ändern, wobei die Bevorzugung der Partei x kontinuierlich

*Abbildung 1.2: Beispiele von Veränderungsprozessen*



zu- oder abnimmt. Zeitdiskrete Änderungen ergeben sich eigentlich nur dann, wenn man untersucht, wie sich diese latente Variable zu den verschiedenen Wahlterminen manifestiert, die bei den Bundestagswahlen bekanntlich im 4-Jahres-Rhythmus stattfinden (vgl. Abb. 1.2a).

Ein ähnliches Beispiel ist die Veränderung der Arbeitszeit. Auch hier könnte man zunächst argumentieren, daß täglich geringfügige Schwankungen der individuellen Arbeitszeit auftreten und allenfalls im Mittel eine gewisse Konstanz zu beobachten wäre. Nur bei Betrachtung tariflich oder betrieblich vereinbarter Arbeitszeiten sind Veränderungen zu diskreten Zeitpunkten zu beobachten, da die entsprechenden Regelungen innerhalb festgelegter Zeitintervalle zwischen den zuständigen Gremien ausgehandelt werden (vgl. Abb. 1.2c).

Man kann daher zunächst einmal festhalten, daß Veränderungen in vielen Fällen an keinen festen Zeitrahmen gebunden sind. Das gilt besonders für Veränderungen auf der Konstruktebene, die im engeren Sinne Gegenstand sozialwissenschaftlicher Theoriebildung ist. *Zeitdiskrete Prozesse* lassen sich in einigen Fällen auf der Ebene der beobachteten (manifesten) Variablen annehmen, wenn organisatorische, institutionelle oder rechtliche Regelungen nur bestimmte Veränderungszeitpunkte zulassen. Der Umkehrschluß, daß es sich bei beobachtbaren Merkmalen immer um einen zeitdiskreten Prozeß handeln muß, ist jedoch nicht zulässig. Das zeigen die anderen drei Beispiele.

Am deutlichsten ist das bei der Körpergröße: Bis zu einem bestimmten Lebensalter nimmt die Körpergröße eines Menschen beständig zu. Auch wenn diese Veränderung innerhalb sehr kleiner Zeitintervalle kaum meßbar ist oder das Körperwachstum manchmal sprunghaft verläuft, dürfte doch jedermann einleuchten, daß in diesem Fall zu jedem Zeitpunkt Veränderungen auftreten. Es ergibt sich eine kontinuierlich zunehmende Wachstumskurve (vgl. Abb. 1.2e).

Bei den beiden anderen Beispielen "Beruf" und "Einkommen" ist die Situation nicht ganz so deutlich (vgl. Abb. 1.2b,d). Auch hier könnte man einwenden, daß beobachtbare Einkommens- und Berufsänderungen an bestimmte Fristen und Termine (z.B. Kündigungsfristen) gebunden sind. Auf der Ebene der manifesten Variablen könnte man also einen zeitdiskreten Prozeß unterstellen. Aus praktischen Gründen erscheint mir dieses Vorgehen aber nicht besonders sinnvoll, denn erstens wären diese Intervalle so klein, daß sie das Modell unnötig komplizieren würden, und zweitens gelten diese Fristen nicht gleichermaßen für alle Untersuchungspersonen.

Der zusätzliche Aufwand steht daher in keinem Verhältnis zu seinem Nutzen. Aus praktischen wie aus theoretischen Gründen halte ich daher die Annahme für sinnvoll, daß Einkommens- und Berufsveränderungen jederzeit auftreten können. In beiden Fällen handelt es sich also wie bei der Körpergröße um einen *zeitkontinuierlichen Prozeß.*

Die Annahme zeitkontinuierlicher Veränderungen ist für die Analyse vieler sozialer Wandlungsprozesse die angemessenere Alternative und dementsprechend sollte die Zeitdimension des Datenwürfels möglichst alle Zeitpunkte des Veränderungsprozesses erfassen. Unter Modellierungsgesichtspunkten ist noch die Frage interessant, *wie* sich die einzelnen Variablen verändern können. In diesem Zusammenhang spielt das Meßniveau eine Rolle: Nicht-metrische Merkmale, wie etwa "Beruf" oder "Zweitstimme", können sich per definitionem nur *diskontinuierlich* ändern, da eine Veränderung immer den Wechsel zwischen zwei qualitativ verschiedenen Zuständen impliziert. Einen solchen abrupten Zustandswechsel bezeichnet man auch als *Ereignis.*

Metrische Merkmale können sich dagegen diskontinuierlich oder *kontinuierlich* ändern, wie die Beispiele "Körpergröße" und "Einkommen" zeigen: Während bei der Körpergröße kontinuierlich Veränderungen auftreten, sind beim Einkommen jederzeit Veränderungen möglich. Erst wenn eine Einkommensänderung eintritt, ist zu entscheiden, wie hoch das Einkommen steigt oder fällt. Modelle für diskontinuierliche Veränderungsprozesse müssen daher zwei Aspekte formalisieren: erstens den Zeitpunkt und zweitens die Art der Veränderung. Ein Modell für kontinuierliche Veränderungsprozesse betrachtet dagegen nur einen Aspekt: Größenänderungen pro Zeiteinheit. Die folgende Tabelle 1.2 faßt die formalen Aspekte der fünf Beispiele noch einmal zusammen.

*Tabelle 1.2: Klassifikation von Veränderungsprozessen*

| Art der Veränderung | Zeitpunkt der Veränderung | |
|---|---|---|
| | diskret | kontinuierlich |
| diskontinuierlich | Zweitstimme[a)]<br>tarifliche Arbeitszeit[b)] | Beruf[a)]<br>Einkommen[b)] |
| kontinuierlich | | Körpergröße[b)] |

Meßniveau: a) nicht-metrisch, b) metrisch

Es liegt auf der Hand, daß Prozesse, die aus einer Folge von Ereignissen (diskontinuierlichen Änderungen) ohne festgelegte Zeitpunkte bestehen, nur durch eine genaue Registrierung von Zeitpunkt und Art der einzelnen Veränderungen rekonstruiert werden können. Die resultierende Datenstruktur bezeichne ich als *Verlaufsdaten*. Sie enthalten exakte Angaben über Art, Abfolge und Zeitpunkte der einzelnen Veränderungen oder Ereignisse und werden daher auch als *Ereignisdaten* bezeichnet. Bei Verlaufsdaten erfaßt die Zeitdimension des Datenwürfels prinzipiell unendlich viele Zeitpunkte, aus praktischen Gründen werden jedoch nur die Zeitpunkte festgehalten, zu denen tatsächlich Änderungen eintreten.[1]

So wünschenswert Verlaufsdaten unter theoretischen Gesichtspunkten sein mögen, häufig bestehen Längsschnittdaten jedoch nur aus zeitdiskreten Messungen. Dafür sprechen zum einen erhebungspraktische Erwägungen (vgl. den folgenden Abschnitt), zum anderen aber auch die prinzipielle Unmöglichkeit, kontinuierliche Veränderungsprozesse, die sich nicht in Form von Ereignissen mit eindeutigen Veränderungszeitpunkten niederschlagen, auch fortlaufend zu protokollieren. Man denke etwa an das Beispiel "Körpergröße": Die kontinuierliche Messung des Größenwachstums eines Kindes verbietet sich von selbst. Von daher wird man die Wachstumskurve des Kindes durch mehrere zeitdiskrete Messungen (näherungsweise) erfassen. Ähnliche Überlegungen gelten natürlich auch für viele Einstellungsfragen der Umfrageforschung: Man kann wohl die Parteipräferenz zu bestimmten Stichtagen feststellen (z.B. in Form der bekannten "Sonntagsfrage"). Eine Frage nach dem Zeitpunkt der letzten Einstellungsänderung ("Wann hat sich Ihre Präferenz für Partei x geändert?") könnte aber wahrscheinlich von den wenigsten Befragten exakt beantwortet werden. Werden also verschiedene Eigenschaften (Variablen) einer Stichprobe von Untersuchungseinheiten zu mehreren diskreten Zeitpunkten erfaßt, dann spricht man von *Paneldaten*. Diese Datenstruktur entspricht dem obigen Datenwürfel N·V·T mit diskreter Zeitachse T.[2]

---

1) Da Anzahl und Zeitpunkte der Änderungen für jede Untersuchungseinheit verschieden sein können, variiert Anzahl und (zeitliche) Verteilung der Informationen für jede Beobachtung und das Bild eines rechtwinkligen Würfels ist nicht mehr ganz angemessen.

2) Zeitreihen unterscheiden sich von Paneldaten dadurch, daß sie üblicherweise nicht zwischen verschiedenen Untersuchungseinheiten unterscheiden, so daß sich eine zweidimensionale Datenmatrix V·T ergibt (s. oben). Außerdem ist die Anzahl der erfaßten Zeitpunkte T in der Regel sehr viel größer als bei Paneldaten. Gleichwohl kann man Paneldaten als N Zeitreihen V·T kürzerer Dauer oder als T Querschnitte N·V betrachten (sogenannte "gepoolte" Querschnitts– und Zeitreihendaten).

Mit dieser Art von Längsschnittdaten wird relativ pragmatisch umgegangen. Im übertragenen Sinne könnte man sagen, man behandelt die ursprünglich dreidimensionale Datenmatrix N·V·T wie eine zweidimensionale Datenmatrix N·VT, die sich von Querschnittsdaten N·V nur dadurch unterscheidet, daß einige Variablen Meßwiederholungen aufweisen. Eine typische Untersuchungsfrage lautet: Was ist der Zustand der Person i zum Zeitpunkt t+1, wenn zum Zeitpunkt t, t−1, t−2, ... bestimmte Informationen über die Person i bekannt sind. Dazu muß man lediglich die Meßwiederholungen (genauer gesagt: die dadurch entstehende serielle Korrelation) durch geeignete statistische Verfahren kontrollieren, dann kann man die aus Querschnittsanalysen bekannten Methoden anwenden (multiple Regression, Logit− und Probitmodelle, Tabellenanalyse; vgl. die Lehrbücher von KESSLER/GREENBERG 1981, HSIAO 1986, PLEWIS 1985, HAGENAARS 1990, ARMINGER/MÜLLER 1990).

Dieser Ansatz ist von verschiedenen Autoren kritisiert worden, vor allem von J. COLEMAN: "In fact, methods of statistics based on static models of association have been applied to the study of change. But because they are based on static models, the approach is both ad hoc and more complicated than necessary. For quantitative research techniques, a failure to base the study of change upon the mathematical tools appropriate to it can lead only to confusion and unnecessary complications" (1968: 431). Man könnte auch sagen, zeitdiskrete Methoden sind eher datenorientiert und weniger modellorientiert (ANDRESS 1984b). Neben diesem prinzipiellen Einwand, der wahrscheinlich nur Puristen überzeugt, haben zeitdiskrete Modelle jedoch vor allem zwei praktische Nachteile: Erstens hängen die geschätzten Effekte von der Wahl der Zeitintervalle ab und zweitens sind Prognosen nicht für jeden Zeitpunkt möglich, sondern nur für ganze Vielfache der ursprünglichen Zeitintervalle.

Die multivariate Analyse von Ereignis− oder Verlaufsdaten umfaßt dagegen Methoden zur Modellierung diskontinuierlicher Veränderungsprozesse, bei denen Veränderungen jederzeit stattfinden können. Auch wenn sich die Darstellung dieser Modellklasse für den Anwender nicht wesentlich von der Darstellung anderer statistischer Verfahren unterscheidet, kann man jedoch ihre grundlegenden statistischen Konzepte (vor allem die Rate) mit den von Coleman genannten "mathematical tools" verbinden, aus denen sich dann verschiedene Implikationen des Prozesses ableiten lassen, wie TUMA und HANNAN (1984: 79ff.) in eindringlicher Form vorführen. Verlaufsdatenanalyse ist daher eher modellorientiert. Da es sich bei diskon-

tinuierlichen Veränderungsprozessen häufig um nicht-metrische Merkmale handelt (vgl. Tabelle 1.2), zumindest aber um Merkmale mit einer begrenzten Anzahl von Kategorien, bezeichnet man Verlaufsdatenanalyse auch als dynamische Analyse kategorialer Daten.

## *1.3 Erhebung von Verlaufsdaten*

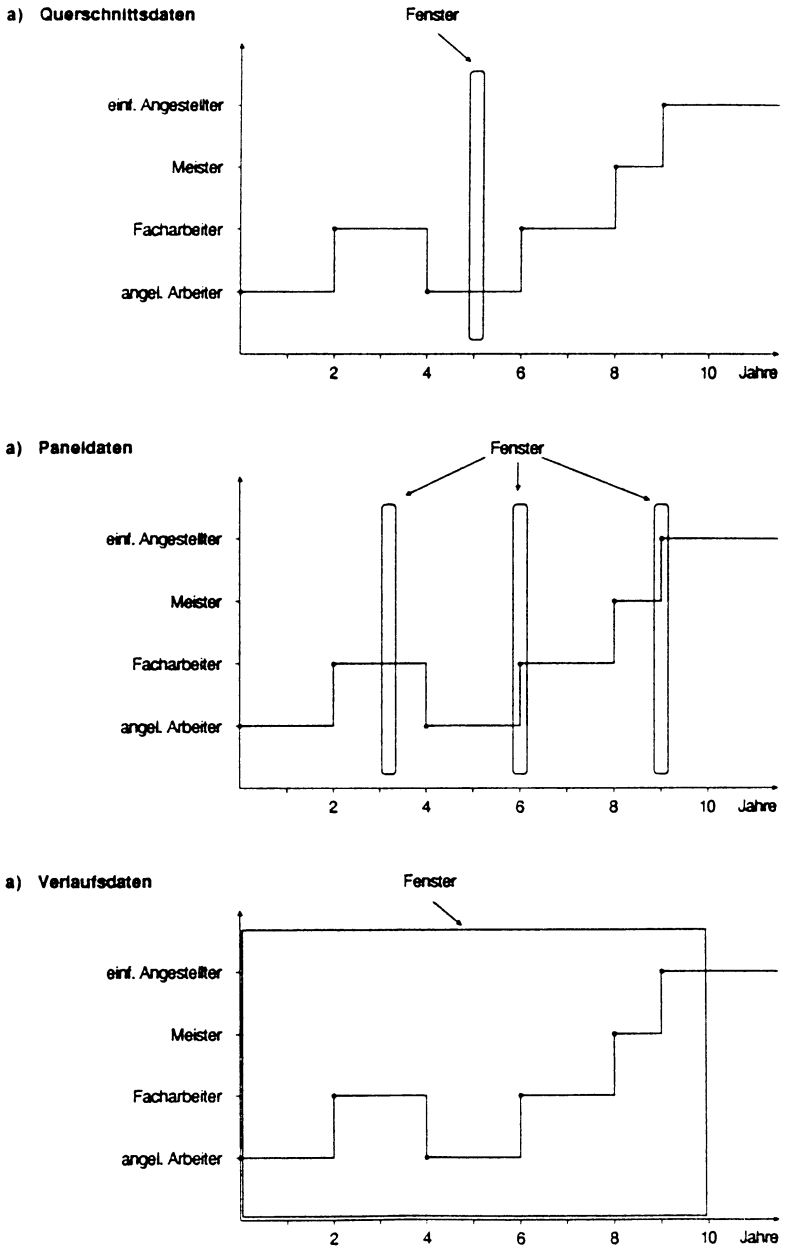
Die weite Verbreitung zeitdiskreter Datenstrukturen deutete es bereits an: Die Erhebung von Verlaufsdaten ist mit sehr viel Aufwand und Kosten verbunden, so daß man häufig auf weniger exakte Daten zurückgreifen muß. Dieser Abschnitt soll zwei Fragen eingehender diskutieren: Wie kann eine möglichst genaue Erhebung zeitkontinuierlicher Verlaufsprozesse durchgeführt werden und warum sind Querschnitts- oder Paneldaten nur ein unzureichender Ersatz? Falls eine solche vollständige Erhebung nicht möglich ist, mit welchen Datenkonstellationen ist man dann in der Forschungspraxis typischerweise konfrontiert?

### *1.3.1 Datenerhebung als Selektionsprozeß*

Der Erhebungsprozeß kann als ein Selektionsprozeß interpretiert werden, in dem der Forscher entweder auf Grund spezifischer Interessen oder auf Grund praktischer Beschränkungen nur einen Teil der vorhandenen Informationen erfaßt. Bildlich gesprochen ist die Erhebung ein Fenster, das über den tatsächlich stattfindenden Verlauf gelegt wird und bestimmte Aspekte des Veränderungsgeschehens sichtbar macht. Abbildung 1.3 illustriert diesen Vergleich an Hand des Berufsverlaufs von Herrn Müller.

In den meisten sozialwissenschaftlichen Untersuchungen werden Veränderungen im Zeitablauf schlichtweg ignoriert oder für unbedeutend gehalten. Mit einer *Querschnittserhebung* wird der Zustand aller Untersuchungseinheiten zu einem bestimmten Zeitpunkt erfaßt (vgl. Abb. 1.3a). Es liegt auf der Hand, daß mit dieser Informationsbasis, wenn überhaupt, nur sehr eingeschränkte Aussagen über Veränderungsprozesse möglich sind.

*Abbildung 1.3: Querschnitts–, Panel– und Verlaufsdaten*



Will man hingegen Veränderungen im Zeitablauf untersuchen, dann stehen einem mehr oder weniger differenzierte Erhebungsverfahren zur Verfügung. Da eine kontinuierliche Erhebung von Veränderungsprozessen mit sehr viel Aufwand und Kosten verbunden ist, werden häufig *zeitdiskrete Erhebungsverfahren* bevorzugt. Sie erfassen den Zustand jeder Untersuchungseinheit zu mindestens zwei verschiedenen Zeitpunkten. Die Anzahl der Erhebungszeitpunkte ist in der Regel begrenzt, insbesondere wenn eine große Anzahl von Einheiten untersucht werden soll. Zeitdiskrete Erhebungsdesigns lassen sich als eine Serie von schmalen Fenstern veranschaulichen, die über den untersuchten Veränderungsprozeß gelegt werden (vgl. Abb. 1.3b). Natürlich ist der "Einblick" bei diesem Beobachtungsverfahren sehr viel größer als bei Querschnittsdaten, Teile der einzelnen Verläufe werden jedoch "übersehen", wie z.B. Herrn Müllers Tätigkeit als angelernter Arbeiter im Baugewerbe. Ganz allgemein läßt sich zeigen, daß bei einem zeitdiskreten Beobachtungsraster Zustände kürzerer Dauer unterrepräsentiert sind (FELLER 1971: 13, SÖRENSEN 1977: 213ff.).

Eine vollständige Erfassung aller Veränderungen ist nur möglich, wenn das zu untersuchende Merkmal *kontinuierlich* über einen längeren Zeitraum *erhoben* wird. Mit anderen Worten, es wird ein großes Fenster über den Veränderungsprozeß gelegt, das innerhalb eines bestimmten Bereiches alle Veränderungen sichtbar macht (vgl. Abb. 1.3c). Ein solches Beobachtungsverfahren ist natürlich sehr aufwendig und kann häufig nur dadurch umgesetzt werden, daß man die untersuchten Personen bittet, eine möglichst genaue retrospektive Schilderung aller Veränderungen zu geben. Dennoch ist auch dieses sehr differenzierte Erhebungsdesign unvollständig, da jeder Erhebungszeitraum zeitlich begrenzt ist. Man kann keine Aussagen darüber machen, wann der (innerhalb des Untersuchungszeitraums) erste Zustand begonnen hat und wann der (innerhalb des Untersuchungszeitraums) letzte Zustand enden wird. Vergangenheit und Zukunft des Prozesses bleiben also unbekannt, weil durch die Begrenzung der Untersuchungsperiode Teile der Verläufe abgeschnitten werden.

Man spricht auch von *links*– bzw. *rechtszensierten Beobachtungen.* Erhebungstechnisch sind Rechtszensuren häufig ein prinzipielles Problem ("man kann nicht in die Zukunft schauen"), während Linkszensuren in einigen Fällen durch (retrospektive) Nacherhebung korrigiert werden können. Unter statistischen Gesichtspunkten sind allerdings Linkszensuren sehr viel schwieriger zu handhaben als Rechtszensuren. Man kann sogar sagen, daß durch geeignete statistische Modellierung das Problem rechtszensierter Beobachtungen weitgehend eliminiert werden kann. Bei historischen Daten

kann das Zensurproblem manchmal vernachlässigt werden, weil der Prozeß von Anfang an verfolgt werden kann und so weit zurückliegt, daß auch rechtszensierte Beobachtungen nicht vorkommen.

Zensierungen sind ein Spezifikum von Ereignisdaten und können verschiedene Ursachen haben. In Abbildung 1.3 treten zensierte Beobachtungen auf Grund eines begrenzten Untersuchungszeitraumes auf. Wenn diese Beobachtungsdauer $\tau$ Zeiteinheiten beträgt, dann hängt die Anzahl der beobachteten Ereignisse davon ab, mit welcher Rate Ereignisse auftreten. In diesem Fall, den man auch als *Typ I Zensierung* bezeichnet, ist die Beobachtungsdauer festgelegt und die *Ereignishäufigkeit* zufällig.[1]

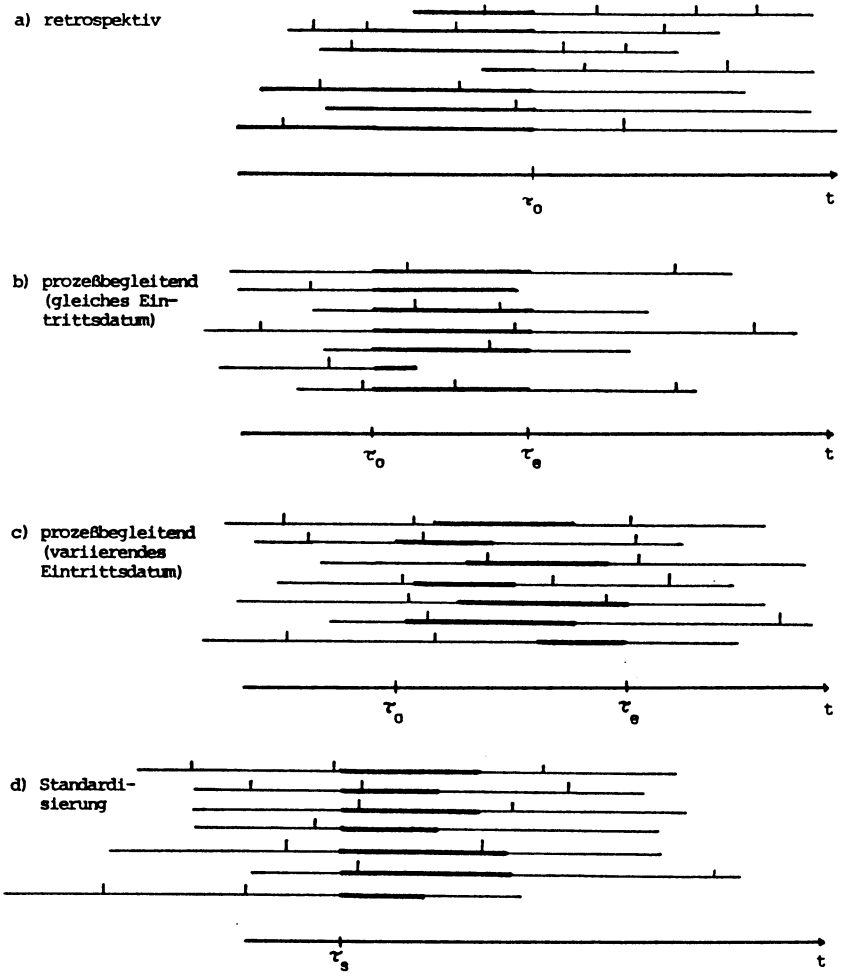
Dies ist jedoch nicht die einzige Möglichkeit. *Typ II zensierte Beobachtungen* ergeben sich, wenn die Anzahl der Ereignisse im vorhinein festgelegt wird. Angenommen ein Experiment wird solange durchgeführt, bis D von insgesamt N Untersuchungseinheiten ein Ereignis aufweisen. In diesem Fall ist die *Wartezeit*, bis das D−te Ereignis auftritt, zufällig. Natürlich hängt auch sie davon ab, mit welcher Rate Ereignisse auftreten. Typ II zensierte Beobachtungen sind statistisch einfacher zu handhaben als Typ I zensierte Beobachtungen (vgl. Abschnitt 5.2.1). Da aber dieser Typ vor allem bei experimentellen Designs mit technischen oder biometrischen Fragestellungen vorkommt, werde ich ihn weitgehend vernachlässigen.

### *1.3.2 Kontinuierliche Erhebung von Verlaufsdaten*

Die folgende Abbildung 1.4 zeigt verschiedene Möglichkeiten, Verlaufsdaten kontinuierlich zu erheben. Dabei ist für jeweils mehrere Untersuchungseinheiten der zeitliche Verlauf eines Merkmals durch eine horizontale Gerade dargestellt. Alle diese Verläufe sind an Hand einer einheitlichen Zeitachse verortet. Die Art der einzelnen Veränderungen wird nicht dargestellt, lediglich der Zeitpunkt ist durch entsprechende Symbole gekennzeichnet. Der Teil des jeweiligen zeitlichen Verlaufs, der durch das Erhebungsdesign (Fenster) erfaßt (sichtbar) wird, ist hervorgehoben gezeichnet. Im einzelnen unterscheide ich folgende Möglichkeiten der Erhebung von Verlaufsdaten:

1) Wie die Diskussion des folgenden Abschnitts zeigen wird, ist die Beobachtungsdauer $\tau$ nicht notwendigerweise für alle Untersuchungseinheiten gleich groß. Dieser und allgemeinere Zensierungsmechanismen werden in Abschnitt 2.5.2 ausführlicher besprochen.

*Abbildung 1.4: Erhebung von Verlaufsdaten*



a) *Retrospektive Erhebung* (vgl. Abb. 1.4a): Von einem Erhebungszeitpunkt $\tau_0$ wird der zeitliche Verlauf so weit es geht zurückverfolgt. Dabei werden alle auftretenden Ereignisse registriert. Dies entspricht dem Erhebungsdesign einer Event–History–Analyse (z.B. die Lebensverlaufsstudie, vgl. BLOSSFELD et al. 1986).

b) *Prozeßbegleitende Erhebung mit gleichem Eintrittsdatum* aller Untersuchungseinheiten (vgl. Abb. 1.4b): Von einem Erhebungszeitpunkt $\tau_0$ werden alle innerhalb des Zeitraumes $\tau_0$ bis $\tau_e$ auftretenden Ereignisse registriert. Das kann innerhalb eines modifizierten Paneldesigns geschehen, bei dem durch retrospektive Erhebungen die Ereignisse zwischen den Panelwellen erfaßt werden (z.B. das sozio–ökonomische Panel, vgl. HANEFELD 1984).

c) *Prozeßbegleitende Erhebung mit variierenden Eintrittsdaten* der Untersuchungseinheiten (vgl. Abb. 1.4c): Innerhalb eines Untersuchungszeitraumes $\tau_0$ bis $\tau_e$ werden die zeitlichen Verläufe der ausgewählten Untersuchungseinheiten aufgezeichnet. Eine solche Situation tritt sehr oft bei prozeßproduzierten oder bei Archivdaten auf: Z.B. in einer medizinischen Untersuchung, bei der die Patientenkarrieren der Personen ausgewertet werden, die innerhalb eines bestimmten Zeitraumes zur stationären Behandlung in ein Krankenhaus eingeliefert werden. Oder aber bei der Analyse der Mandatsdauern von Abgeordneten, die innerhalb des Kaiserreiches Mitglied des Deutschen Reichstages waren (ANDRESS et al. 1992).

Die drei Möglichkeiten unterscheiden sich in mehreren Aspekten. Zu nennen wären die Ziehung der Stichprobe, die Anzahl der Erhebungskontakte, die Aktualität der erhobenen Information und die Länge der aufgezeichneten zeitlichen Verläufe. Ich will diese Unterschiede kurz erläutern und dann ihre Konsequenzen für die Auswertung von Verlaufsdaten besprechen.

Bei den Alternativen a und b wird die Stichprobe zu einem bestimmten Termin $\tau_0$ gezogen, während bei Alternative c die Auswahl der Erhebungseinheiten über den gesamten Untersuchungszeitraum $\tau_0$ bis $\tau_e$ verteilt ist. Bei Alternative a findet dann eine einmalige Erhebung statt, während in den beiden anderen Fällen mehrere Erhebungen durchgeführt werden. Hier werden also in der Tat kontinuierlich Informationen gesammelt. Dementsprechend aktuell sind diese Daten, während bei Alternative a Erinnerungsverluste in Kauf genommen werden müssen. Schließlich variiert in allen drei Fällen die Länge der erhobenen Verläufe. Bei Alternative c ist das

ganz besonders deutlich, denn hier ist die unterschiedlich lange Beteiligung der Untersuchungseinheiten eine Folge variierender Eintrittsdaten (*progressive Zensierung*).

Je nachdem, in welcher Richtung die Aufzeichnung der einzelnen Verläufe erfolgt, ist man mit spezifischen Erhebungsproblemen konfrontiert. Bei prozeßbegleitenden Erhebungen sind das im wesentlichen Ausfälle von Untersuchungspersonen, Gewöhnungseffekte sowie Veränderungen der Forschungsziele und Erhebungsinstrumente im Fortgang der Untersuchung. Bei Retrospektiverhebungen überwiegen vor allem Erinnerungsprobleme. Darüber hinaus sind ältere Personen auf Grund unterschiedlicher Mortalität unterrepräsentiert. Eine ausführliche Diskussion dieser Erhebungsprobleme findet sich in ANDRESS (1984b).

Bei Alternative a und b erfolgt die Aufzeichnung der einzelnen Verläufe für alle Untersuchungseinheiten während des gleichen Zeitraums beginnend jeweils mit dem Datum $\tau_0$. Eine solche Untersuchungsgruppe, die durch einen gemeinsamen Anfangspunkt definiert ist, bezeichnet man auch als *Kohorte* (engl. *cohort data*). Zeiteffekte (hier insbesondere Effekte des Erhebungsdesigns) wirken auf alle Untersuchungseinheiten in gleicher Weise.

Das gilt nicht unbedingt für die dritte Möglichkeit (engl. *follow–up data*). Nur unter der *Annahme konstanter Rahmenbedingungen* lassen sich die zeitlich unterschiedlich verorteten Verläufe an Hand des Eintrittsdatums $\tau_s$ in die Studie synchronisieren (vgl. Abb. 1.4d). Je nachdem, wann eine Untersuchungseinheit in die Studie aufgenommen wurde, kann man einen mehr oder weniger langen Verlauf beobachten.

Dagegen sind bei Kohortendaten häufig alle Verläufe gleich lang. Aus praktischen Gründen sind aber auch hier Abweichungen möglich, wenn z.B. bei prozeßbegleitender Erhebung einzelne Untersuchungspersonen frühzeitig ausscheiden oder bei retrospektiver Erhebung jüngere Personen prinzipiell nicht länger zurückverfolgt werden können. Möchte man schließlich verschiedene Altersgruppen unterschiedlicher Kohorten miteinander vergleichen, dann ist bei allen drei Alternativen schon ein sehr aufwendiges Design notwendig, um für jede Gruppe genügend zeitbezogene Informationen zu haben (zur Analyse von Alters–, Kohorten– und Periodeneffekten vgl. HAGENAARS 1990, GLENN 1977).

### *1.3.3 Ungenaue Erhebung von Verlaufsdaten*

So wünschenswert es auch sein mag, Veränderungsprozesse möglichst vollständig zu erfassen, finanzielle, materielle, personelle und zeitliche Restriktionen verhindern häufig eine extensive Anwendung kontinuierlicher Erhebungsverfahren. In vielen Untersuchungen muß man daher mit gruppierten, aggregierten oder zeitdiskret erhobenen Verlaufsdaten vorliebnehmen:

a) *gruppierte Verlaufsdaten*: Erinnern wir uns noch einmal an Herrn Müller (vgl. Tabelle 1.1). Angenommen man hätte ihn gebeten, für die Untersuchung einen beruflichen Lebenslauf zu schreiben, in dem alle seine bisherigen Tätigkeiten mit genauen Anfangs- und Enddaten aufgeführt werden. Diese Aufgabe wäre Herrn Müller sicherlich schwergefallen, insbesondere wenn er sich an sehr kurze und weit zurückliegende Tätigkeiten erinnern sollte. Also wird man Herrn Müller bitten, einen tabellarischen Lebenslauf zu erstellen, der nur jeweils eine, die "hauptsächliche" Tätigkeit pro Jahr nennt. Dabei sollen Tätigkeiten unter sechs Monaten Dauer unberücksichtigt bleiben. Abgesehen von der Vernachlässigung kurzer Tätigkeiten handelt es sich offensichtlich um gruppierte Verlaufsdaten, da bei Erhebung in Jahresintervallen nicht mehr erkennbar ist, wann innerhalb des jeweiligen Jahres eine Tätigkeit begann.

b) *zeitdiskret erhobene Verlaufsdaten*: Natürlich kann man den Erhebungsaufwand auch dadurch reduzieren, daß man den beruflichen Lebenslauf nur zu bestimmten Stichtagen (z.B. Ende 1964, 1967, 1970) erhebt.

c) *aggregierte Verlaufsdaten*: Schließlich hätte man Herrn Müller auch fragen können, wie oft er in den letzten 10 Jahren seinen Beruf gewechselt hat. Dadurch wird die Vielzahl von Verlaufsinformationen in einer Kennzahl zusammengefaßt (aggregiert).

Betrachtet man den Erhebungsprozeß im Anschluß an die obige Diskussion als Selektionsprozeß, dann läßt sich Alternative a quasi als zeitraumbezogene und Alternative b als zeitpunktbezogene Auswahl aus den vorliegenden Verlaufsinformationen charakterisieren ("interval" und "point sampling" bei CHAMBERLAIN 1984). Beide Erhebungsverfahren sind insofern ungenau, als sie nicht alle Veränderungen abbilden, die tatsächlich stattgefunden haben. Dies gilt in besonderem Maße für Alternative c, in der die Variation entlang der Zeitdimension in einer Zahl verdichtet wird. In diesem Fall ist es kaum noch möglich, Veränderungen des Prozesses im Zeitablauf zu testen. Beschränken wir daher die Diskussion auf die Alternativen a und b.

An dieser Stelle muß man nun entscheiden, ob man weiterhin an einem zeitkontinuierlichen Prozeß festhalten möchte oder ob man angesichts der Datenlage gleich einen *zeitdiskreten Prozeß* modelliert. Dieses Vorgehen ist zwar theoretisch nicht besonders befriedigend (s.o.), erfordert aber weniger Annahmen bei der folgenden Auswertung. Möchte man dennoch an einem *zeitkontinuierlichen Prozeß* festhalten, dann benötigt man weitere Annahmen über Art und Zeitpunkt der ungenau erfaßten Veränderungen. Die einfachste Lösung wäre, die Ungenauigkeiten zu ignorieren und anzunehmen, daß die sich ergebenden Meßfehler die Auswertung nicht negativ beeinflussen. Also interpretiert man die ungenau erhobenen Veränderungen als Ereignisse und verwendet als Veränderungszeitpunkte ersatzweise das Stichjahr oder die Mitte des Erhebungsintervalls. Je nachdem, wie groß die Ungenauigkeiten sind, ist dieses Vorgehen natürlich nicht besonders befriedigend.

Untersucht man einen Prozeß, in dem zwar verschiedene Ereignisse möglich sind, die sich aber nicht wiederholen können, dann ist dieses Vorgehen noch einigermaßen tragbar. Man muß lediglich Annahmen über den Zeitpunkt des Ereignisses machen, während die Art des Ereignisses eindeutig feststeht. Angenommen Personen dürfen nur einmal ihren Beruf wechseln. Eine Person sei zu Beginn einer Längsschnittuntersuchung als Facharbeiter tätig. Zu einem späteren Zeitpunkt gibt sie an, als Angestellter beschäftigt zu sein. Egal zu welchem Zeitpunkt diese Angabe gemacht wurde, es läßt sich eindeutig feststellen, daß ein Wechsel vom Facharbeiter zum einfachen Angestellten stattgefunden hat. Lediglich der Zeitpunkt dieses Wechsels ist unbekannt.

Anders stellt sich die Situation dar, wenn, wie in der Realität, mehrere Berufswechsel möglich sind. Unter dieser Voraussetzung kann die befragte Person natürlich auf ganz unterschiedlichen Wegen Angestellter geworden sein. Je weiter die beiden Angaben auseinanderliegen, um so größer ist die Wahrscheinlichkeit, daß sich in der Zwischenzeit weitere Berufswechsel ereignet haben. In dieser Situation ist es hilfreich, zumindest partielle Informationen über die vergangenen Veränderungen zu erhalten (z.B. Anzahl Wechsel, Art und Zeitpunkt des letzten Wechsels etc.).

### *1.3.4 Zusammenfassung*

Zusammenfassend kann man feststellen, daß Verlaufsdaten möglichst exakt erhoben werden sollten. Wenn das der Fall ist, dann ist für jede Untersuchungseinheit Abfolge, Art und Zeitpunkt aller im Untersuchungszeitraum stattfindenden Ereignisse bekannt. Je nachdem, wie viele Ereignisse im Untersuchungszeitraum auftreten, variiert die Anzahl der Informationen pro Untersuchungseinheit.

Auf Grund praktischer Restriktionen ist es jedoch in vielen Fällen nicht möglich, eine solche differenzierte Datenbasis zu erheben. Die Information steht dann in mehr oder weniger ungenauer Form zur Verfügung. Das sind entweder Informationen über den Zustand der Untersuchungseinheiten zu verschiedenen diskreten Zeitpunkten oder Angaben über repräsentative Zustände innerhalb festgelegter Erhebungsintervalle. Pro Untersuchungseinheit gibt es eine feste Anzahl von Informationen, da alle dem gleichen Beobachtungsraster unterworfen sind. Durch Vergleich aufeinanderfolgender Zustände erhält man gewisse Aufschlüsse über die real stattfindenden Veränderungen, jedoch können einzelne Ereignisse unberücksichtigt bleiben.

*Tabelle 1.3: Temporale Datenstrukturen*

a) zeitkontinuierliche Verlaufsdaten

| Person | Ereignis Nr. | Zeitpunkt des Ereignisses | Art des Ereignisses | Ausgangs– zustand | Zielzustand |
|---|---|---|---|---|---|
| Müller | 1 | 2 | Aufstieg | angel. Arbeiter | Facharbeiter |
| Müller | 2 | 4 | Abstieg | Facharbeiter | angel. Arbeiter |
| Müller | 3 | 6 | Aufstieg | angel. Arbeiter | Facharbeiter |
| Müller | 4 | 8 | Aufstieg | Facharbeiter | Meister |
| Müller | 5 | 9 | Aufstieg | Meister | einf. Angestellter |

b) zeitdiskrete Verlaufsdaten

| Person | Erhebung Nr. | Zeitpunkt der Erhebung | Art des Ereignisses | Zustand zum Zeitpunkt der Erhebung |
|---|---|---|---|---|
| Müller | 1 | 3 | – | Facharbeiter |
| Müller | 2 | 6 | kein Wechsel | Facharbeiter |
| Müller | 3 | 9 | Aufstieg | einfacher Angestellter |

Beide Möglichkeiten sind an Hand des fiktiven Berufsverlaufs von Herrn Müller in Tabelle 1.3 dargestellt. Im ersten Fall wurde angenommen, daß alle stattfindenden Veränderungen exakt erhoben wurden. Im zweiten Fall wurde eine zeitdiskrete Erhebung im Abstand von jeweils 3 Jahren unterstellt (z.B. durch ein Panel, vgl. Abbildung 1.3b). Man sieht noch einmal sehr deutlich, wie ungenau der Berufsverlauf im zweiten Fall abgebildet wird. Im ersten Fall variiert die Anzahl der Datensätze pro Untersuchungseinheit, während im zweiten Fall diese Zahl immer gleich ist. Man beachte außerdem, daß andere Formen ungenauer Erhebung ebenfalls zeitdiskrete Daten ergeben: Statt der Erhebungszeitpunkte gibt es dann Erhebungsintervalle.

Ich bezeichne die erste Alternative als *zeitkontinuierliche Daten*, weil hier jeder Punkt auf der Zeitskala definiert ist. Dagegen können bei der zweiten Alternative nur bestimmte Werte auftreten. Man könnte sogar die Zeitangaben ganz weglassen, denn sie ergeben sich aus der Abfolge der Erhebungen und der Kenntnis der Erhebungsintervalle. Ich bezeichne diese Datenstruktur daher als *zeitdiskrete Daten*. Zeitkontinuierliche Daten sind natürlich die theoretisch bessere Alternative und die exaktere Informationsbasis. Da man jedoch häufig mit zeitdiskreten Daten konfrontiert ist, müssen zeitkontinuierliche Modelle, wenn möglich, an diese Datenstruktur angepaßt werden.

# 2. *Statistische Grundlagen der Verlaufsdatenanalyse*

Der *zeitliche Verlauf* eines Merkmals Z(t) beschreibt die Sequenz der Merkmalsausprägungen (Zustände), die im Zeitablauf bei einer Untersuchungseinheit auftreten. Jeder zeitliche Verlauf besteht aus einer Folge von *Ereignissen (Zustandsänderungen, Wechseln, Übergängen)*. Dahinter verbergen sich eine Fülle von Detailinformationen:

- In welcher Reihenfolge treten die einzelnen Zustände/Ereignisse auf?
- Welche Dauer haben die einzelnen Zustände bzw. zu welchen Zeitpunkten finden die einzelnen Zustandsänderungen (Ereignisse) statt?
- Zwischen welchen Zuständen tritt ein Übergang auf bzw. um welches Ereignis handelt es sich?

In der Regel wird man nicht nur eine sondern mehrere Untersuchungseinheiten betrachten. Bei allen sind Veränderungen möglich, man kann jedoch nicht genau angeben, wann diese Ereignisse auftreten werden und um welche Ereignisse es sich dabei handeln wird. Alle Untersuchungseinheiten sind also einem Ereignisrisiko ausgesetzt, das stochastischer Natur ist. Man bezeichnet sie daher auch als *Risikomenge*. Empirisch feststellbar ist allein eine *Verteilung von Zustandsdauern* (*Wartezeiten*) und *Übergängen* zwischen verschiedenen Zuständen. Die hier zu besprechenden Modelle sollen diese Daten so beschreiben, daß es möglich ist, einzelne Aspekte eines Verlaufsprozesses zu prognostizieren und zu erklären.

Dazu bedient man sich der *Theorie stochastischer Prozesse*. Da dieses Spezialgebiet der mathematischen Statistik die unterschiedlichsten Prozesse betrachtet, ist es für diese anwendungsbezogene Darstellung notwendig, eine gewisse Auswahl der Themen vorzunehmen. In Abschnitt 2.1 erläutere ich dazu einige Grundbegriffe, die im Laufe der folgenden Abhandlung häufiger verwendet werden und die es vor allem gestatten, eine gewisse Ausgrenzung von Themen vorzunehmen, die im Rahmen dieser Abhandlung nicht interessieren. Wie man sich weiter vorstellen kann, ist es angesichts der oben beschriebenen Datenfülle notwendig, zunächst mit ganz einfachen stochastischen Prozessen zu beginnen und diese dann schrittweise zu verallgemeinern.

Stochastische Prozesse kann man unter den verschiedensten Problemstellungen diskutieren. In diesem Kapitel möchte ich vor allem die Frage untersuchen, ob es eine zentrale Variable für stochastische Prozesse gibt,

aus der sich alle anderen Ergebnisse ableiten lassen. Es zeigt sich, daß die *Rate*, mit der Ereignisse auftreten, ein solcher zentraler Parameter ist. Neben der Erläuterung der wesentlichen statistischen Begriffe und der Darstellung stochastischer Prozesse unterschiedlicher Komplexität soll in diesem Kapitel gezeigt werden, welche Aspekte dieser Prozesse mit Hilfe von Raten prognostiziert werden können. Wie man diese Modelle dann mit Hilfe empirischer Daten überprüft (schätzt), ist Gegenstand des empirischen Teils dieser Arbeit (insb. Kapitel 4 und 5).

## *2.1 Grundbegriffe und Eingrenzung des Themengebiets*

Eine erste Klassifikation stochastischer Prozesse ergibt sich, wenn man den betrachteten Zustandsraum sowie Art und zeitliches Auftreten der einzelnen Ereignisse näher analysiert. Handelt es sich bei dem untersuchten Merkmal Z(t) z.B. um den Beruf einer Person, der durch ein Erhebungsschema mit 20 Berufskategorien erfaßt wird, dann ist die Zahl der möglichen Zustände, die eine Person im Zeitablauf einnehmen kann, begrenzt. Man spricht von einem *endlichen Zustandsraum*. Wird statt dem jeweiligen Beruf die Zahl der Arbeitsplatzwechsel gemessen, dann ist diese Zahl prinzipiell unbegrenzt, wenn auch praktisch nicht unendlich viele Wechsel auftreten. In diesem Fall ist der *Zustandsraum abzählbar unendlich*. Schließlich kann auch der Berufsstatus mit einer kontinuierlichen Statusskala erhoben worden sein, so daß ein *kontinuierlicher Zustandsraum* vorliegt. Von der Definition des Zustandsraumes hängt die Gestaltung der Ordinate eines *Verlaufsdiagramms* ab (vgl. Abb. 1.2).

Eine andere Klassifikationsdimension betrifft die Zeitachse dieses Verlaufsdiagramms. Je nachdem, ob Veränderungen nur zu diskreten Zeitpunkten oder jederzeit stattfinden können, sind nur bestimmte Punkte der Zeitachse t definiert. Man unterscheidet daher stochastische Prozesse mit diskreten oder stetigen Zeitparametern oder einfach *Prozesse mit stetiger oder diskreter Zeit*.

Ich beschränke mich im folgenden auf stochastische Prozesse mit endlichem Zustandsraum und stetigen Zeitparametern. Anders ausgedrückt, ich betrachte die Veränderung kategorialer Variablen und gehe dabei davon aus, daß Veränderungen jederzeit stattfinden können. Diese Beschränkung der hier zu betrachtenden Modelle rechtfertigt sich durch die Tatsache, daß

die meisten sozialen Prozesse kontinuierlich verlaufen und viele sozialwissenschaftliche Merkmale nicht-metrisches Meßniveau haben. Davon unabhängig ist die Frage, wie der jeweilige Veränderungsprozeß erhoben wurde. Wie in Abschnitt 1.3.3 diskutiert, kann das sehr ungenau geschehen. Man muß daher in einigen Fällen berücksichtigen, daß die Daten zeitdiskret erhoben wurden, obwohl der eigentliche Prozeß kontinuierlich verläuft.

Mit der Definition des Zustandsraums ergibt sich die Frage, welche Beziehungen zwischen den einzelnen Zuständen bestehen und welche Ereignisse möglich sind. Ein Zustand, der nicht verlassen werden kann, ist ein *absorbierender Zustand.* Das jeweilige Ereignis, d.h. der Beginn dieses Zustands, kann sich daher nicht wiederholen. Beispiele wären etwa das Ausscheiden aus einer Paneluntersuchung oder das Ende der ersten Tätigkeit, wenn nur der Erstberuf interessiert (vgl. Datei A in Abschnitt 1.1.3). In vielen Fällen tritt ein Zustand jedoch mehrmals im Zeitablauf auf: Jemand ist beispielsweise in verschiedenen Wirtschaftszweigen als angelernter Arbeiter beschäftigt. Tätigkeitswechsel wiederholen sich also im Zeitablauf (*wiederholbare Ereignisse*; vgl. auch Datei C) und dementsprechend unterscheidet man zwischen mehreren aufeinanderfolgenden Tätigkeiten, die man auch als *Episoden* (engl. spells) bezeichnet. Weiterhin ist es sinnvoll, danach zu unterscheiden, ob nur ein bestimmtes oder mehrere Ereignisse möglich sind. Man spricht von *singulären bzw. multiplen Ereignissen.* Singuläre Ereignisse sind häufig nicht wiederholbar[1]. Zuverlässigkeitsprüfungen in der industriellen Fertigung konzentrieren sich z.B. auf ein bestimmtes Ereignis, den Verschleiß des untersuchten Bauteils, während in einer soziologischen Studie der Karrieremobilität mehrere Ereignisse, d.h. verschiedene Arten von Berufswechseln untersucht werden (vgl. Datei B). Multiple, aber nicht wiederholbare Ereignisse werden in medizinischen Anwendungen auch *konkurrierende Risiken* genannt.

Wie die Diskussion der verschiedenen Auswertungsstrategien zeigte (vgl. Abschnitt 1.1.3), hat die Frage, ob man nun singuläre, multiple und/oder wiederholbare Ereignisse untersucht, häufig weniger mit der Natur des untersuchten Prozesses als mit den jeweiligen Forschungsinteressen zu tun. Die Definition des Zustandsraumes ist daher eine Forscherentscheidung

1) Der Zustandsraum besteht praktisch aus zwei Zuständen. Einer davon ist ein absorbierender Zustand.

und dementsprechend sollte man von *Modellen* sprechen. BLOSSFELD et al. (1986) bevorzugen daher die Unterscheidung in Ein– und Mehr–Zustands– bzw. Ein– und Mehr–Episoden–Modelle.

In der Praxis untersucht man nicht nur einzelne zeitliche Verläufe sondern Daten einer größeren Stichprobe von Untersuchungseinheiten. Je nachdem, ob sich die Zahl der Untersuchungseinheiten im Zeitablauf verändert oder gleich bleibt, unterscheidet man zwischen *offenen und geschlossenen Systemen.* Im ersteren Fall muß man neben dem eigentlichen stochastischen Prozeß auch den Zu– und Abgang von Untersuchungseinheiten kontrollieren.

Ich will mich auf die Analyse geschlossener Systeme konzentrieren, also auf die Frage, wie eine gegebene Untersuchungsgruppe im Zeitablauf den Zustandsraum "durchwandert". Der stochastische Prozeß beginnt mit einer *Anfangsverteilung* der Untersuchungseinheiten auf die einzelnen Zustände zum Zeitpunkt $t_0$ und bei optimaler Erfassung des Veränderungsprozesses läßt sich der *Systemzustand* (Verteilung der Einheiten auf die einzelnen Zustände) zu jedem Zeitpunkt t angeben. Interessant ist dabei die langfristige Entwicklung des Systems, insbesondere die Frage, ob sich das System auf einen stabilen *Gleichgewichtszustand* einpendelt.

Wenn man eine Gruppe von Untersuchungseinheiten über einen längeren Zeitraum beobachtet, ergeben sich zwei weitere Unterscheidungsmerkmale stochastischer Prozesse. Einmal ist es wahrscheinlich, daß nicht für alle Untersuchungseinheiten der gleiche Veränderungsprozeß gilt. Man unterscheidet daher *homogene und heterogene Populationen.* Zum anderen ist es plausibel, daß sich die Gesetzmäßigkeiten stochastischer Prozesse selbst im Zeitablauf verändern. Der Prozeß selbst ist zeitabhängig. Je nachdem, spricht man von *zeitkonstanten bzw. stationären oder zeitveränderlichen bzw. nicht–stationären Prozessen.*

Auf zwei weitere Möglichkeiten der Klassifikation stochastischer Prozesse möchte ich an dieser Stelle nur kurz eingehen. Sie betreffen einmal die Zahl der betrachteten zeitveränderlichen Merkmale und zum anderen das verwendete Meßmodell. Wenn man mehrere Merkmale im Zeitverlauf erhebt, dann spricht man von einem *multidimensionalen Zustandsraum.* Man hat die Möglichkeit, durch geeignete statistische Operationen die Veränderung bei einem Merkmal durch Veränderungen bei anderen Merkmalen zu erklären.

*Meßfehler* haben bei zeitbezogenen Daten eine besondere Bedeutung. Bei der Diskussion verschiedener Erhebungsverfahren habe ich schon darauf hingewiesen, daß Messungen so ungenau sein können, daß man sozusagen Veränderungen übersieht (vgl. Abschnitt 1.3.1). Es ist aber auch die umgekehrte Situation denkbar: Man produziert durch fehlerhafte Messungen künstlich Veränderungen, wo gar keine sind. Wahrscheinlich ist auch, daß man Fehler im Zeitablauf wiederholt, so daß mehrere aufeinanderfolgende Beobachtungen scheinbar voneinander abhängig sind. Alle praktisch anwendbaren Methoden der Verlaufsdatenanalyse berücksichtigen zum gegenwärtigen Zeitpunkt nur in eingeschränktem Maße das Vorhandensein von Meßfehlern.

Im Zusammenhang mit der Erhebung von Verlaufsdaten sei auch noch einmal an die spezifischen Datenprobleme erinnert: In den wenigsten Fällen ist es möglich, einen stochastischen Prozeß vollständig von Anfang bis Ende zu beobachten. Einige Zustandsdauern werden daher nur unvollständig erfaßt. Insbesondere kann man keine Aussagen über die folgenden Zustandsänderungen (Ereignisse) machen. Teilweise ist der Beobachtungszeitraum sogar so kurz, daß einzelne Untersuchungseinheiten überhaupt kein Ereignis aufweisen. Man spricht hier von *Zensierung* und je nachdem, ob der Anfang oder das Ende eines Prozesses unvollständig erfaßt wurde, unterscheidet man *links- bzw. rechtszensierte Beobachtungen.*

Die folgende Darstellung erfolgt schrittweise. Eine hilfreiche begriffliche Differenzierung liefert dazu LANCASTER (1990: xi): Er unterscheidet zwischen "duration" und "transition data". Die frühen Arbeiten aus der eher technisch orientierten Zuverlässigkeitsprüfung oder der biometrisch orientierten Survival Analysis beschäftigen sich mit dem Eintritt eines bestimmten Ereignisses (Ausfall einer Maschine, Tod einer Untersuchungsperson). Dementsprechend steht die Analyse der *Zeitdauer* bis zum Eintritt des Ereignisses (Zufallsvariable T) im Vordergrund der statistischen Analyse.[1] Ich beginne daher meine Einführung in Abschnitt 2.2 und 2.3 mit solchen einfachen Prozessen, in denen nur ein singuläres, nicht-wiederholbares Ereignis auftritt. Während Abschnitt 2.2 eher die Funktion hat, die grundlegenden Konzepte der Verlaufsdatenanalyse zu verdeutlichen, geht es in Abschnitt 2.3 um einen Überblick über die Vielzahl der Verteilungsmodelle für Wartezeiten.

1) Für diesen Anwendungsfall sind im übrigen die meisten Auswertungsprogramme zur Verlaufsdatenanalyse konzipiert – zumindest die der großen statistischen Programmpakete.

Wie die Beispiele in der Einführung zeigen, sind Verlaufsprozesse in den Sozialwissenschaften jedoch in der Regel komplexer. Es treten Veränderungen zwischen mehreren Zuständen auf und dies nicht nur einmal, sondern mehrmals im Zeitablauf. Dementsprechend sind hier neben der Zeitdauer auch die verschiedenen *Übergänge* (ggfs. mit Wiederholung) Gegenstand der statistischen Analyse. Neben der Zufallsvariablen T ist jetzt auch noch die Zufallsvariable Z(t) zu betrachten. Es können jedoch einige der Erkenntnisse aus einfacheren Prozessen übertragen werden, wenn man das Datenmaterial geeignet aggregiert (z.B. alle Episoden eines Ausgangszustands zusammenfaßt). Abschnitt 2.4 beschäftigt sich mit einigen dieser komplexeren Prozesse und stellt die statistischen Konzepte zur Analyse multipler und/oder wiederholbarer Ereignisse soweit vor, wie sie für die empirischen Auswertungen in Kapitel 4 und 5 benötigt werden.

Die Abhandlung der statistischen Grundlagen schließt in Abschnitt 2.5 mit zwei Komplikationen der zuvor besprochenen Prozesse. Zum einen geht es um die Frage heterogener Verlaufsprozesse, in denen nicht alle Untersuchungseinheiten das gleiche Veränderungsverhalten zeigen. Zum anderen sollen die Konsequenzen unvollständiger Beobachtung (Zensierung) diskutiert werden.

## *2.2 Prozesse mit singulären, nicht – wiederholbaren Ereignissen*

Angenommen man will untersuchen, wann Erwerbstätige nach Berufseintritt das erste Mal ihren Beruf wechseln. Man verwendet Datei A und zählt, wieviel Personen pro Jahr ihre erste Tätigkeit beenden. Das Ergebnis dieser Auswertung für 256 zufällig ausgesuchte Personen zeigt Tabelle 2.1. Dabei treten keine zensierten Beobachtungen auf. Dies erleichtert den Einstieg in die folgenden Ableitungen.

In diesem Abschnitt möchte ich mich mit der Frage beschäftigen, mit welchen Maßzahlen man diese Daten beschreiben kann. Aus didaktischen Gründen verwende ich dazu zunächst zeitdiskrete Daten. Die entsprechenden Maßzahlen für kontinuierliche Daten kann man sich dann am besten dadurch vorstellen, daß die Erhebungsintervalle unendlich klein werden. Um die Betrachtung nicht zu sehr zu komplizieren, werde ich bei den zeitdiskreten Maßzahlen die Dauer der Erhebungsintervalle ignorieren. Ich

gehe davon aus, daß alle Ereignisse jeweils am Ende eines Jahres auftreten, unterstelle also einen *zeitdiskreten Prozeß.*[1]

*Tabelle 2.1: Häufigkeit erster Tätigkeitswechsel nach Jahren*

| Jahr | Personen mit Wechsel | Personen ohne Wechsel | Anteil Wechsel | kumulativer Anteil Wechsel | Anteil an Personen ohne Wechsel | Anteil ohne Wechsel |
|---|---|---|---|---|---|---|
| $t_k$ | $d_k$ | $n_k$ | $\hat{f}(t_k)$ | $\hat{F}(t_k)$ | $\hat{q}_k$ | $\hat{S}(t_k)$ |
| 1 | 167 | 256 | 0,652 | 0,652 | 0,652 | 1,000 |
| 2 | 48 | 89 | 0,188 | 0,840 | 0,539 | 0,348 |
| 3 | 23 | 41 | 0,090 | 0,930 | 0,561 | 0,160 |
| 4 | 6 | 18 | 0,023 | 0,953 | 0,333 | 0,070 |
| 5 | 3 | 12 | 0,012 | 0,965 | 0,250 | 0,047 |
| 6 | 6 | 9 | 0,023 | 0,988 | 0,667 | 0,035 |
| 7 | 1 | 3 | 0,004 | 0,992 | 0,333 | 0,012 |
| 8 | 1 | 2 | 0,004 | 0,996 | 0,500 | 0,008 |
| 9 | 1 | 1 | 0,004 | 1,000 | 1,000 | 0,004 |

### *2.2.1 Wartezeitverteilungen*

Die Häufigkeiten in Spalte 2 der Tabelle sind eine *empirische Wartezeitverteilung.* Es geht um die Zeit bis zum ersten Tätigkeitswechsel oder anders ausgedrückt, um die Dauer der ersten Tätigkeiten. Wie man sieht, werden sehr viele Tätigkeiten nicht länger als ein Jahr ausgeübt und nur wenige Personen sind länger als 5 Jahre in ihrer ersten Tätigkeit beschäftigt. In Abbildung 2.1 ist diese Wartezeitverteilung graphisch dargestellt. Sie ist rechtsschief. Diese Asymmetrie beobachtet man sehr häufig bei Wartezeitverteilungen, da sich viele Ereignisse entweder zu Beginn oder (seltener) am Ende eines Prozesses häufen.

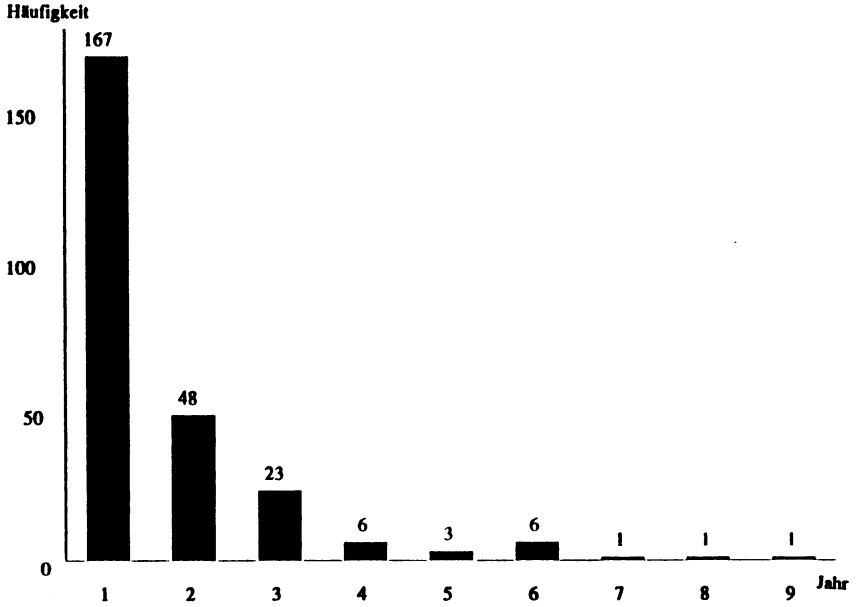
1) Eine Alternative wären gruppierte Wartezeiten. Man benötigt dann jedoch eine Annahme darüber, wie sich die ursprünglich zeitkontinuierlichen Wartezeiten über das Intervall verteilen. Diese etwas kompliziertere Situation werden wir erst in Kapitel 4 behandeln. Für diese Einführung beschränken wir uns auf den einfachen Fall, daß sich alle Ereignisse auf einen Zeitpunkt, das Intervallende, konzentrieren.

Die Daten lassen sich auf unterschiedliche Art und Weise beschreiben:

1. Zunächst einmal kann man sich fragen, in welchem Jahr Tätigkeitswechsel am wahrscheinlichsten sind.
2. Dann kann man untersuchen, wie wahrscheinlich es ist, daß eine Person länger als 5 Jahre im ersten Beruf tätig ist.
3. Da sich möglicherweise die Mobilitätsrate im Zeitablauf verändert, interessiert schließlich, wie groß das Risiko eines Wechsels in jedem Jahr ist, vorausgesetzt es hat bis dato noch kein Wechsel stattgefunden.

Könnte man schließlich ein statistisches Verteilungsmodell angeben, dann wäre die Beschreibung der Daten besonders einfach, denn ein solches Modell erlaubt es bekanntlich, alle Aspekte einer Verteilung zu berechnen.

*Abbildung 2.1: Eine empirische Wartezeitverteilung*




Um die drei Fragen beantworten zu können, verwendet man verschiedene charakteristische Funktionen, die jeweils unterschiedliche Aspekte der Verteilung charakterisieren. Es handelt sich dabei um

1. die *Sterbewahrscheinlichkeiten* (bei diskreten Wartezeiten) bzw. die *Dichtefunktion* (bei kontinuierlichen Wartezeiten – death density function),
2. die *Verteilungsfunktion* der *Überlebenswahrscheinlichkeiten* (survivor function) und
3. die *Risikofunktion* oder *Rate* (hazard function).

An der Namensgebung erkennt man die Anleihen aus medizinischen Anwendungen. Da die meisten Standardtexte zur Survival Analysis englischsprachig sind, habe ich die Originalbezeichnungen in Klammern angefügt. Ich finde die Übersetzung nicht in allen Fällen besonders glücklich. Mangels besserer Alternativen habe ich sie dennoch übernommen, auch wenn sich für spezifische Anwendungen bessere Bezeichnungen finden lassen (s. unten).

In den folgenden Abschnitten möchte ich die drei Funktionen definieren und im zeitdiskreten Fall an Hand der Beispieldaten aus Tabelle 2.1 illustrieren. Da alle drei Funktionen verschiedene Aspekte desselben Prozesses beschreiben, liegt es auf der Hand, daß sie nicht unabhängig voneinander sind. In einem weiteren Abschnitt möchte ich daher die Zusammenhänge zwischen den drei Funktionen diskutieren. Sterbe– und Überlebenswahrscheinlichkeiten lassen sich nämlich beide auf die Rate zurückführen. Schließlich möchte ich zeigen, wie man Verteilungsmodelle für Wartezeiten entwickelt, indem man bestimmte Annahmen über den zeitlichen Verlauf der Rate macht. Eine allgemeine Übersicht über Verteilungsmodelle für Wartezeiten folgt dann in Abschnitt 2.3.

### *2.2.2 Diskrete Wartezeiten*

Angenommen die Zufallsvariable T (Tätigkeitsdauer) kann nur diskrete Zahlwerte $t_k$ ($k=1,2,...$) annehmen ($T\in\mathbb{N}$). Tätigkeitswechsel können nur zu diesen diskreten Zeitpunkten stattfinden, dazwischen ist die Wahrscheinlichkeit eines Wechsels Null.

Die Funktion $f(t_k)$, die sogenannte *Sterbewahrscheinlichkeit*, mißt nun die Wahrscheinlichkeit, daß ein Ereignis zum Zeitpunkt $T=t_k$ stattfindet:

$$f(t_k) = P(T=t_k) \tag{2.1a}$$

Für die Daten in Tabelle 2.1 ist der Anteil aller Ereignisse $d_k$ pro Intervall bezogen auf den Stichprobenumfang N ein unverzerrter Schätzer dieser Wahrscheinlichkeit:

$$\hat{f}(t_k) = \frac{d_k}{N} \tag{2.2}$$

Spalte 4 der Tabelle zeigt diese Anteilswerte für jedes Jahr. Man erkennt, daß Tätigkeitswechsel im ersten Jahr am wahrscheinlichsten sind. In diesem Anwendungsfall hätte man die Funktion $f(t_k)$ besser Wechsel– als Sterbewahrscheinlichkeit genannt. Offensichtlich handelt es sich um eine *Wahrscheinlichkeitsfunktion*, wie sie für die Analyse diskreter Zufallsvariablen aus der schließenden Statistik bekannt ist (vgl. den mathematischen Anhang). Dementsprechend hat sie die folgenden Eigenschaften:

1. Wie alle Wahrscheinlichkeiten kann sie nur positive Werte im Bereich [0,1] annehmen.

$$0 \leq f(t_k) \leq 1 \qquad \text{für } k = 1,2,... \tag{2.3a}$$

2. Die Summe aller Einzelwahrscheinlichkeiten ergibt 1.

$$\sum_k f(t_k) = 1 \tag{2.4a}$$

Die Funktion $S(t_k)$, die sogenannte *Überlebensfunktion*, mißt die Wahrscheinlichkeit, den Zeitpunkt $t_k$ zu erleben:

$$S(t_k) = P(T \geq t_k) \tag{2.5a}$$

Betrachten wir dazu Tabelle 2.1. Die Summe aller Wechsel (Ereignisse) ergibt den Stichprobenumfang N. Das ist zugleich die Anzahl der Personen, die zu Beginn des Prozesses dem Ereignisrisiko ausgesetzt ist. Nach dem ersten Jahr sind 167 Personen mit einem Tätigkeitswechsel ausgeschieden und es verbleiben 89 Personen, die noch ihre Tätigkeit wechseln können. Der Anteil dieser *Risikomenge* $n_k$ (risk set) am Stichprobenumfang N ist ein unverzerrter Schätzer für die diskrete Überlebenswahrscheinlichkeit zum jeweiligen Zeitpunkt:

(2.6a) $$\hat{S}(t_k) = \frac{n_k}{N}$$

Auch diese Werte sind in der Tabelle (Spalte 7) berechnet. Man erkennt, wie der Anteil der Personen ohne Wechsel stufenweise abnimmt – wieviel Prozent z.B. die ersten 5 Berufsjahre ohne Tätigkeitswechsel überstehen. Durch Inspektion der Überlebenswahrscheinlichkeit $S(t_k)$ kann man also angeben, wieviel Prozent einer Untersuchungsgruppe einen bestimmten Zeitpunkt überleben. Man bezeichnet $S(t_k)$ daher auch als kumulativen *Überlebensanteil* oder als *Überlebensfunktion*. Es lassen sich damit auch Aussagen über die mittlere Lebenserwartung einer Untersuchungsgruppe machen, in unserem Fall also die mittlere Dauer des ersten Berufs.

Die *Überlebenswahrscheinlichkeit* $S(t_k)$ läßt sich auch direkt aus dem kumulativen Anteil der Personen mit Wechsel $F(t_k)$ berechnen. Man bezeichnet $S(t_k)$ daher auch als *komplementäre Verteilungsfunktion*: Während die Verteilungsfunktion eine monoton steigende Funktion mit Ausgangswert 0 und Endwert 1 ist, ist $S(t_k)$ eine monoton fallende Funktion mit Anfangswert 1 und Endwert 0 (zum Konzept der Verteilungsfunktion vgl. ebenfalls den mathematischen Anhang). Konkret erhält man den kumulativen Überlebensanteil $S(t_k)$, indem man den kumulierten Anteil der Personen mit Wechsel zum vorherigen Zeitpunkt $t_{k-1}$ von 1 subtrahiert (mit $F(t_0) = f(t_0) = 0$):

(2.7a) $$S(t_k) = 1 - P(T < t_k) = 1 - F(t_{k-1}) = 1 - \sum_{i=0}^{k-1} f(t_i)$$

Dieser Rechenvorgang läßt sich auch umkehren. Man erhält die Sterbewahrscheinlichkeit $f(t_k)$, indem man die Differenz zweier aufeinanderfolgender Werte der Überlebenswahrscheinlichkeit bildet:

(2.8a) $$f(t_k) = S(t_k) - S(t_{k+1}) = F(t_k) - F(t_{k-1})$$

Da das Risiko eines Ereignisses möglicherweise im Zeitablauf variiert, ist es sinnvoll, eine Wahrscheinlichkeit zu berechnen, die diese Veränderungen berücksichtigt. Die *bedingte Sterbewahrscheinlichkeit* $q_k$ mißt die Wahrscheinlichkeit, daß ein Individuum zu einem Zeitpunkt $t_k$ ein Ereignis aufweist, vorausgesetzt es hat bis zum Zeitpunkt $t_k$ ohne Ereignis überlebt. Sie ist folgendermaßen definiert:

(2.9a) $$q_k = P(T=t_k | T \geq t_k)$$

Auch hier machen die Daten aus Tabelle 2.1 am besten deutlich, was mit diesem Konzept gemeint ist. Da es hier um die Wahrscheinlichkeit eines Ereignisses geht, vorausgesetzt es hat bis dato noch kein Wechsel stattgefunden, berechnet man jetzt den Anteil der Wechsel $d_k$ pro Zeitpunkt an den noch verbliebenen Personen ohne Wechsel $n_k$ und erhält einen unverzerrten Schätzer für $q_k$:

(2.10) $$\hat{q}_k = \frac{d_k}{n_k}$$

In einem zeitdiskreten Prozeß wird $q_k$ auch als *zeitdiskrete Rate* bezeichnet. Spalte 6 der Tabelle 2.1 enthält die Schätzwerte der zeitdiskreten Rate. Auf Grund der großen Schwankungen ist keine eindeutige Tendenz der Mobilitätsrate erkennbar.

$q_k$ ist nichts anderes als eine bedingte Wahrscheinlichkeit. Sie kann daher Werte zwischen 0 und 1 annehmen. Man erhält sie auch dadurch, daß man die Sterbe– durch die Überlebenswahrscheinlichkeit teilt:

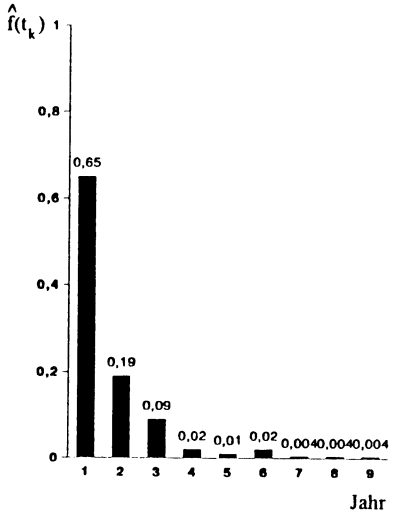
(2.11a) $$q_k = \frac{f(t_k)}{S(t_k)}$$

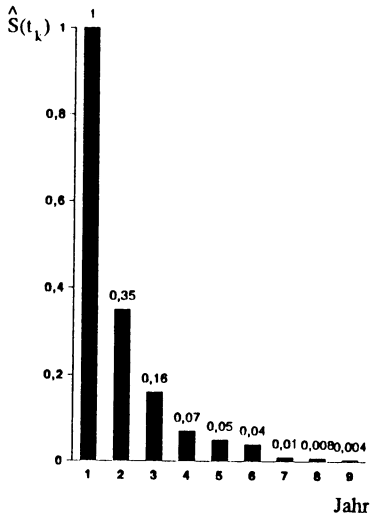
wie man leicht durch Division der entsprechenden Schätzer $\hat{f}(t_k)$ und $\hat{S}(t_k)$ nachrechnen kann. Eine Umformung von (2.11a) zeigt schließlich, daß man die Sterbewahrscheinlichkeit $f(t_k)$ auch durch Multiplikation von $q_k$ und der Überlebenswahrscheinlichkeit $S(t_k)$ erhält. Alle drei Funktionen stehen also miteinander in Verbindung. Abbildung 2.2 zeigt sie noch einmal im Überblick.

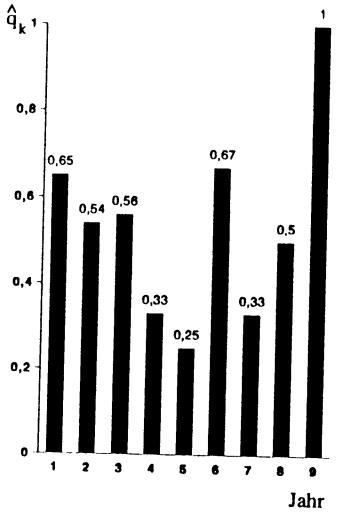
### *2.2.3 Kontinuierliche Wartezeiten*

Für kontinuierliche Zufallsvariablen T lassen sich ähnliche Funktionen f(t), S(t) und r(t) definieren ($T \geq 0$, $T \in \mathbb{R}$). Die mathematischen Ableitungen werden jedoch zwangsläufig für den Nicht–Mathematiker unanschaulicher. Ganz grob kann man sagen, daß das, was wir bisher durch Summierung und Subtraktion erreichen konnten, nun durch Integration und Differentia-

*Abbildung 2.2: Überlebensfunktion, Sterbewahrscheinlichkeit und Rate für zeitdiskrete Daten*



tion erledigen müssen. Ich werde daher f(t), S(t) und r(t) in Analogie zum zeitdiskreten Modell entwickeln und zur besseren Vergleichbarkeit identische Gleichungsnummern verwenden. Allerdings hat dieser Vergleich auch seine Grenzen und wie wir gleich bei der Dichte f(t) sehen werden, können nicht alle Schlußfolgerungen des vorherigen Abschnitts unisono übertragen werden.

Gehen wir also einmal davon aus, daß die Tätigkeitsdauern ursprünglich zeitkontinuierlich gemessen wurden. Eine Maßzahl, die sich ohne Abstriche auf kontinuierliche Zufallsvariablen übertragen läßt, ist die *Überlebenswahrscheinlichkeit.* S(t) mißt auch im kontinuierlichen Fall die Wahrscheinlichkeit, daß ein Individuum den Zeitpunkt t ($t>0$) erlebt:

$$S(t) = P(T \geq t) = 1 - P(T<t) = 1 - F(t) \tag{2.5b}$$

Ein einfacher Schätzer ist der Anteil der Personen $n_t$, die zum Zeitpunkt t noch kein Ereignis hatten:

$$\hat{S}(t) = \frac{n_t}{N} \tag{2.6b}$$

Dieser läßt sich für jeden beliebigen Zeitpunkt t berechnen, vorausgesetzt, es treten keine zensierten Beobachtungen auf (s. dazu Kapitel 4). Im Prinzip handelt es sich dabei um das Komplement $1-F(t)$ der (empirischen) Verteilungsfunktion (vgl. mathematischer Anhang), daher auch der Name *komplementäre Verteilungsfunktion.* S(t) ist wie im zeitdiskreten Fall eine monoton fallende Funktion mit Anfangswert 1und Endwert 0.

Für die Darstellung in Tabelle 2.1 seien nun die (kontinuierlichen) Tätigkeitsdauern in Jahresintervalle klassifiziert worden. $P(t \leq T<t+\Delta t)$ sei die Wahrscheinlichkeit, daß die Zufallsvariable T in eines dieser Intervalle fällt (Intervalle mit Beginn $t=0,1,\ldots,8$ und Breite $\Delta t=1$). Ein Schätzer dieser Wahrscheinlichkeit ist (2.2). Stellen wir uns weiter vor, daß diese Intervalle immer kleiner werden. Die sogenannte *Dichtefunktion* der kontinuierlichen Wartezeitverteilung ergibt sich dann durch folgende Grenzwertbetrachtung:

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T<t+\Delta t)}{\Delta t} = \frac{dF(t)}{dt} = F'(t) \tag{2.1b}$$

Die Dichte f(t) entspricht der ersten Ableitung $F'(t)=dF(t)/dt$ der Verteilungsfunktion nach der Zeit (vgl. mathematischer Anhang). In der technischen Statistik spricht man auch von *globaler Ausfallrate* f(t). Andere verwenden den Ausdruck *globale Zielereignisdichte* f(t). Eine Dichtefunktion für Wartezeiten hat folgende formale Eigenschaften:

(2.3b)
$$f(t) \geq 0 \quad \text{für } t \geq 0,$$
$$f(t) = 0 \quad \text{für } t < 0,$$

(2.4b)
$$\int_{-\infty}^{\infty} f(t)dt = \int_{0}^{\infty} f(t)dt = 1$$

Im Gegensatz zur Sterbewahrscheinlichkeit (2.1a) kann sie nicht mehr als Wahrscheinlichkeit interpretiert werden.[1] Eine Zeichnung der Funktion f(t) vermittelt jedoch einen optischen Eindruck davon, in welchen Bereichen der Verteilung Ereignisse wahrscheinlicher sind.[2] Man denke etwa an die "Glocken"kurve der Normalverteilungsdichte, die besagt, daß kleine Abweichungen vom Mittelwert der Verteilung wahrscheinlicher sind als große.

Möchte man die Wahrscheinlichkeit $P(t \leq T < t+\Delta t)$ mit Hilfe der Dichte berechnen, dann ist die Fläche unter der Dichtefunktion f(t) im Intervall $[t,t+\Delta t)$ zu bestimmen – also die Funktion f(t) im Bereich $[t,t+\Delta t)$ zu integrieren:

$$P(t \leq T < t+\Delta t) = \int_{t}^{t+\Delta t} f(u)du$$

Da die Verteilungsfunktion F(t) dem Integral der Dichte f(t) entspricht (vgl. mathematischer Anhang), läßt sich diese Wahrscheinlichkeit übrigens sehr viel einfacher auf Grund von F(t) bzw. der Überlebenswahrscheinlichkeit S(t) berechnen:

---

1) Leider läßt sich beim Verweis auf die Dichte f(t) der Begriff "Wahrscheinlichkeit" nicht immer umgehen. Ich werde dann den Ausdruck f(t)dt verwenden, um deutlich zu machen, daß die Wahrscheinlichkeit kontinuierlicher Zufallsvariablen nur für Wertebereiche durch Integration der Dichte zu bestimmen ist.

2) Hohe Werte der Dichte f(t) besagen, daß sich für diese Zeitpunkte die Verteilungsfunktion, d.h. der kumulative Anteil der Personen mit Ereignis, besonders schnell ändert.

$$P(t \leq T < t+\Delta t) = \int_0^{t+\Delta t} f(u)du - \int_0^{t} f(u)du$$
$$= F(t+\Delta t) - F(t) = S(t) - S(t+\Delta t)$$

Allgemein werden dadurch folgende Beziehungen zwischen Dichte und Überlebenswahrscheinlichkeit deutlich:

(2.7b) $$S(t) = 1 - F(t) = 1 - \int_0^{t} f(u)du$$

(2.8b) $$f(t) = \frac{dF(t)}{dt} = \frac{-dS(t)}{dt}$$

Danach ergibt sich die Überlebenswahrscheinlichkeit durch Integration der Dichte und umgekehrt die Dichte durch Differentiation (Ableitung) der Überlebenswahrscheinlichkeit.

Abschließend kann man auch eine zeitkontinuierliche *Rate* definieren, die sich ähnlich wie die Dichte durch Grenzwertbetrachtung aus der bedingten Sterbewahrscheinlichkeit ergibt:

(2.9b) $$r(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t+\Delta t | T \geq t)}{\Delta t}$$

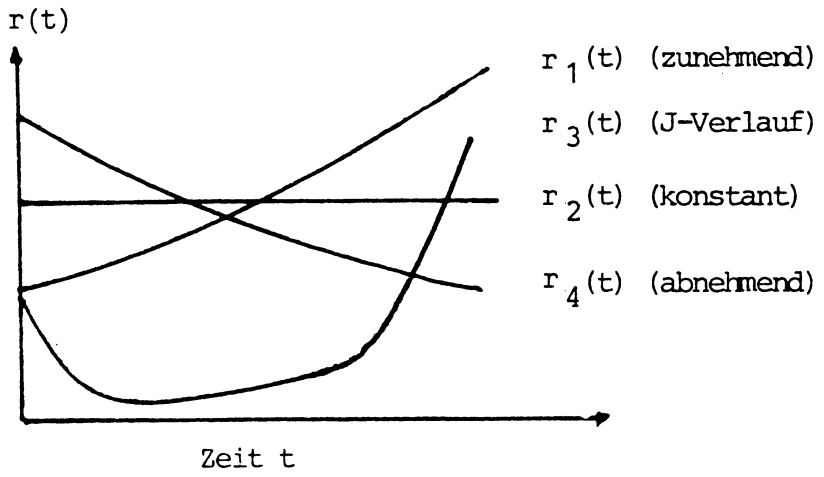
Natürlich handelt es sich auch hier nicht mehr um eine bedingte Wahrscheinlichkeit, d.h. r(t) ist nicht auf das Intervall [0,1] beschränkt und kann jeden Wert $r(t) \geq 0$ annehmen. Die Rate mißt quasi die momentane Neigung zu einem Tätigkeitswechsel, vorausgesetzt die Person hat ihre Tätigkeit bis zum Zeitpunkt t noch nicht verändert. Analog Gleichung (2.11a) entspricht auch die zeitkontinuierliche Rate r(t) dem Quotienten aus Dichte und Überlebenswahrscheinlichkeit:

(2.11b) $$r(t) = \frac{f(t)}{S(t)}$$

In technischen Anwendungen wird r(t) auch als *Ausfalls-* oder *Abgangsrate* bezeichnet. Andere sprechen von *lokaler Zielereignisdichte*, die sich im Gegensatz zur globalen Zielereignisdichte f(t) auf die zeitpunktspezifische

Dynamik des Prozesses bezieht. Für die Daten in Tabelle 2.1 würde sich die Bezeichnung *Mobilitätsrate* anbieten.

*Abbildung 2.3: Idealtypische Verläufe der Rate*




Offensichtlich kann man aus der Rate Veränderungen des Ereignisrisikos im Zeitablauf erkennen. Da das Konzept der Rate außer in den Ingenieurwissenschaften (HARTUNG 1982, STANGE 1970) kaum zur Grundausbildung in der Statistik gehört, sollen in Abbildung 2.3 einige typische Verlaufsprozesse mit ihren entsprechenden Raten vorgestellt werden:

a) Die zunehmende Kurve $r_1(t)$ charakterisiert einen Prozeß, in dem das Ereignisrisiko im Zeitablauf zunimmt. Beispiele: Steigende Mortalitätsrate mit zunehmendem Alter (positive aging) oder erhöhte Störanfälligkeit älterer Maschinen (wear out).

b) Die konstante Kurve $r_2(t)$ charakterisiert einen Prozeß, in dem das Ereignisrisiko zu jedem Zeitpunkt gleich ist. Beispiel: Zufälliges Unfallrisiko.

c) Es sind auch Prozesse wie der J−Verlauf der dritten Rate $r_3(t)$ denkbar, in denen das Ereignisrisiko zunächst abnimmt und dann wieder zunimmt. Beispiel Demographie: Die Mortalitätsrate ist nach der Geburt hoch (Säuglingssterblichkeit), verläuft dann mehr oder weniger konstant auf niedrigem Niveau, um in den letzten Lebensabschnitten wieder zuzunehmen.

d) Der abnehmende Verlauf der vierten Rate $r_4(t)$ könnte schließlich auf unser Mobilitätsbeispiel zutreffen: Je länger jemand in einem Beruf tätig ist, um so geringer ist die Wahrscheinlichkeit eines Wechsels. Andere Beispiele sind der Verlauf der Mortalitätsrate in den ersten Lebensjahren (negative aging) oder das Verhalten von Maschinen in der Installationsphase (burn in).

In allen vier Fällen handelt es sich wohlgemerkt um idealtypische Verläufe. Empirische Schätzungen der Rate werden in der Regel ein sehr viel unregelmäßigeres Bild zeigen. Hat man keine zwingenden theoretischen Gründe für einen bestimmten Ratenverlauf, kann man entweder den Verlauf durch geeignete Funktionen approximieren oder Schätzmethoden verwenden, die keine Annahmen über den zeitlichen Verlauf benötigen, gleichwohl zeitliche Veränderungen zumindest kontrollieren können (vgl. das Regressionsmodell von COX in Abschnitt 5.5). Alle vier Beispiele zeigen jedoch, daß man verschiedene Prozesse eindeutig mit Hilfe der Rate charakterisieren kann. Implizit nehmen auch sozialwissenschaftliche Begriffe auf das Konzept der Rate Bezug (vgl. etwa den Begriff "Mobilitätsrate"). Von daher scheint es sinnvoll, von den drei Funktionen f(t), S(t) und r(t), die eine Wartezeitverteilung charakterisieren, vor allem die Rate r(t) zum Gegenstand statistischer Modellierung zu machen. Diesen Gedankengang möchte ich im folgenden Abschnitt weiterverfolgen.

### *2.2.4 Zusammenhänge der drei Funktionen*

Alle drei Funktionen f(t), S(t) und r(t) charakterisieren unterschiedliche Aspekte einer Wartezeitverteilung, beziehen sich aber auf den gleichen Gegenstand und sind daher nicht unabhängig voneinander. Die Gleichungen (2.7), (2.8) und (2.11) sind Beispiele für Beziehungen der Funktionen untereinander. In diesem Abschnitt möchte ich zeigen, wie sie sich auf die Rate zurückführen lassen.

Ich beginne zunächst mit der Überlebensfunktion S(t) für kontinuierliche Zufallsvariablen. Auf Grund von Gleichung (2.8b) und (2.11b) läßt sich eine Verbindung zwischen der Rate r(t) und S(t) herstellen:

$$r(t) = \frac{f(t)}{S(t)} = \frac{\frac{-dS(t)}{dt}}{S(t)}$$

Man erhält folgende Differentialgleichung für S(t):

$$\frac{dS(t)}{dt} = -r(t)\, S(t) \tag{2.12}$$

Durch Integration erhält man eine Lösung für S(t):

$$S(t) = \exp\left[\int_0^t r(u)du\right] = \exp[-H(t)] \tag{2.13}$$

Man nennt H(t) daher auch die *integrierte* oder *kumulierte Rate*. Die Dichtefunktion f(t) ergibt sich schließlich, indem man S(t) noch einmal mit der Rate multipliziert:

$$f(t) = r(t)\, S(t) = r(t) \exp[-H(t)] \tag{2.14}$$

Dies ergibt sich unmittelbar aus der Definition der Rate (vgl. Gleichung 2.11b).

Ähnliche Beziehungen lassen sich auch für die zeitdiskreten Funktionen entwickeln. Wenn man (2.8a) in (2.11a) einsetzt, ergibt sich

$$q_k = \frac{S(t_k) - S(t_{k+1})}{S(t_k)} = 1 - \frac{S(t_{k+1})}{S(t_k)}$$

Daraus läßt sich eine Rekursionsformel für die zeitdiskrete Überlebenswahrscheinlichkeit $S(t_{k+1}) = S(t_k)(1-q_k)$ ableiten und es ergibt sich allgemein:

$$S(t_k) = \prod_{i=0}^{k-1} (1 - q_i) \quad \text{mit } q_0 = 0 \tag{2.15}$$

Die Sterbewahrscheinlichkeit ergibt sich wiederum durch nochmalige Multiplikation mit der zeitdiskreten Rate:

$$(2.16) \qquad f(t_k) = q_k\, S(t_k) = q_k \prod_{i=0}^{k-1} (1 - q_i) \quad \text{mit } q_0 = 0$$

### 2.2.5 *Implikationen eines bestimmten Verlaufs der Rate für die Verteilung der Wartezeiten*

Nachdem ich gezeigt habe, daß sowohl für kontinuierliche als auch für zeitdiskrete Daten alle charakteristischen Funktionen einer Wartezeitverteilung auf die Rate zurückgeführt werden können, will ich mich jetzt der Verteilung insgesamt zuwenden. Ich beschränke mich dabei auf zeitkontinuierliche Prozesse und möchte zeigen, daß sich bestimmte Verteilungen ergeben, wenn man eine Annahme über den Verlauf der Rate macht.

Ein sehr einfaches Modell ist die Annahme, daß die Rate im Zeitablauf konstant ist: $r(t)=\lambda$. In diesem Fall ist die kumulierte Rate $H(t)=\lambda t$. Durch Einsetzen in (2.13) bzw. (2.14) ergibt sich:

$$(2.17a) \qquad S(t) = \exp(-\lambda t)$$

$$(2.17b) \qquad f(t) = \lambda \exp(-\lambda t)$$

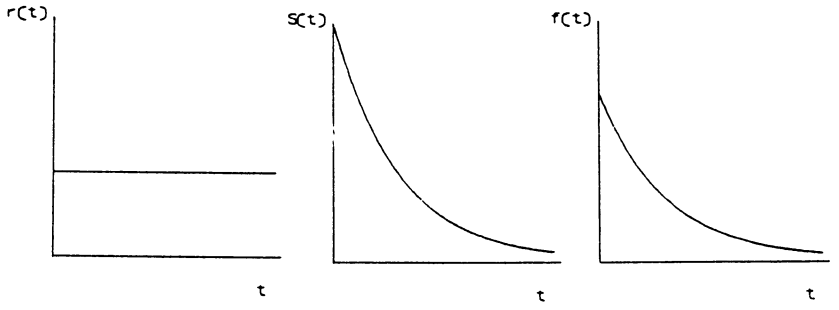
Die Dichte (2.17b) ist auch als *Exponentialverteilung* bekannt. Auf Grund der Annahme eines zeitkonstanten Prozesses ergibt sich also eine ganz bestimmte Verteilung der Wartezeiten. Abbildung 2.4 zeigt Rate, Überlebensfunktion und Dichte der Exponentialverteilung.

Der Vorteil eines Verteilungsmodells besteht bekanntlich darin, daß es die Berechnung verschiedener Aspekte der Verteilung unter den Modellannahmen erlaubt. Der Erwartungswert, der aus der Dichte einer Verteilung berechnet werden kann (vgl. den mathematischen Anhang), gibt z.B. Auskunft über die durchschnittlich zu erwartende Wartezeit.

$$(2.17c) \qquad E(t) = \int_0^\infty t\, f(t)dt = \int_0^\infty t\lambda \exp(-\lambda t)dt = \frac{1}{\lambda}$$

Für eine Exponentialverteilung ergibt sich danach eine Durchschnittsdauer, die dem Kehrwert der Rate entspricht. Bei einer Rate von $\lambda = 0{,}5$ dauert es also durchschnittlich zwei Jahre, bis ein Tätigkeitswechsel eintritt. Dieses Ergebnis legt umgekehrt folgende Interpretation einer (zeitkonstanten) Rate nahe: Eine Rate mißt, wie viele Ereignisse pro Zeiteinheit auftreten (bei $\lambda = 0{,}5$ also 0,5 pro Zeiteinheit oder ein Ereignis pro zwei Zeiteinheiten).

*Abbildung 2.4: Exponentialverteilung*




Aus der Überlebensfunktion kann man den Median oder jedes andere Quantil p der Verteilung berechnen. Z.B. gilt für den Median $\tilde{t}$ mit $p = 0{,}5$

(2.17d) $$p = 0{,}5 = \exp(-\lambda\tilde{t}) \Leftrightarrow \tilde{t} = \ln 2 / \lambda,$$

so daß es bei $\lambda = 0{,}5$ insgesamt 1,39 Jahre dauert, bis 50% der Stichprobe einmal die Tätigkeit gewechselt haben. Andere Quantils–Werte ergeben sich analog mit $-\ln p / \lambda$. Schließlich erkennt man, daß nach $\lambda^{-1}$ Zeiteinheiten $\exp(-1) = 0{,}368$ oder 36,8% der Stichprobe noch keinen Tätigkeitswechsel hatten. Anders ausgedrückt, die durchschnittliche Zustandsdauer ist gleichzeitig der Zeitpunkt, bis zu dem $100 - 36{,}8 = 63{,}2\%$ der Stichprobe die Tätigkeit gewechselt haben.

Die Annahme eines konstanten Ereignisrisikos ist für die folgenden Überlegungen eine sinnvolle Anfangsvoraussetzung. Dieses Modell exponentiell verteilter Wartezeiten fungiert quasi als *Basis–* oder *Vergleichsmodell* für komplexere zeitabhängige Prozesse. Dabei kann man sich für r(t) beliebig komplizierte Funktionen der Zeit einfallen lassen. Durch Integration der Rate (kumulierte Rate) kann man dann jederzeit die Überlebens– und die Dichtefunktion angeben. Einige dieser komplexeren Raten, die in den folgenden Abhandlungen häufiger zitiert werden, sind die

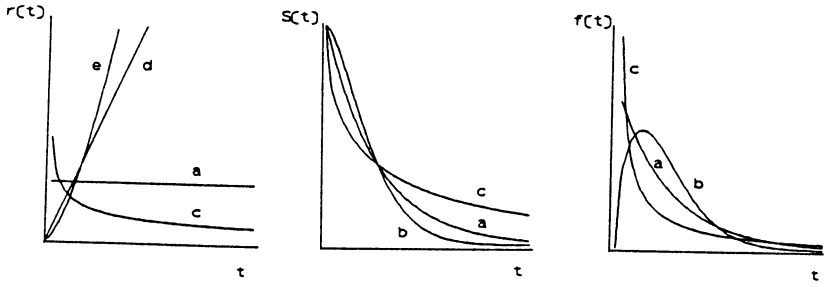
– *Weibull–Rate* $r(t) = \lambda\gamma(\lambda t)^{\gamma-1}$ mit $H(t) = (\lambda t)^{\gamma}$

– *Gompertz–Rate* $r(t) = \lambda\exp(\gamma t)$ mit $H(t) = \frac{\lambda}{\gamma}(\exp(\gamma t)-1)$

– *log–logistische Rate* $r(t) = \frac{\lambda\gamma(\lambda t)^{\gamma-1}}{1+(\lambda t)^{\gamma}}$ mit $H(t) = \ln(1+(\lambda t)^{\gamma})$

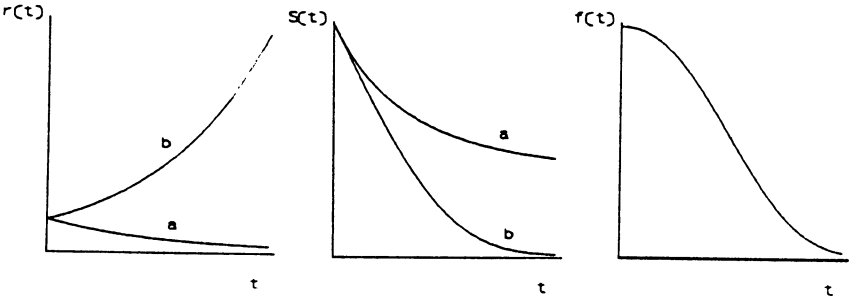
Aus der jeweiligen kumulierten Rate kann man sich mit den Gleichungen (2.13) und (2.14) Überlebensfunktion und Dichte berechnen. Deren Verlauf zeigen die folgenden Abbildungen. Dort ist zu erkennen, daß ein monoton

*Abbildung 2.5: Weibull–Verteilung*



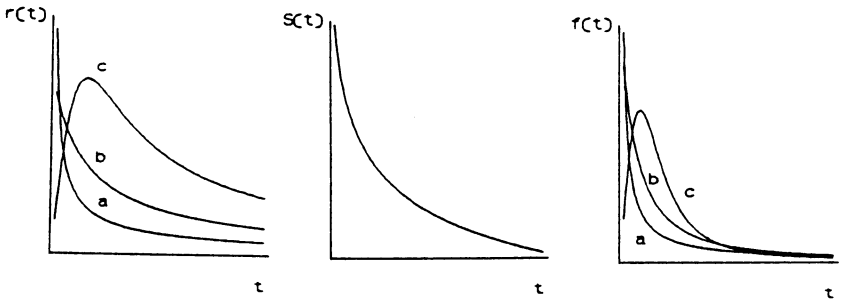

a) $\gamma = 1$, b) $\gamma > 1$, c) $\gamma < 1$, d) $\gamma = 2$, e) $\gamma > 2$

*Abbildung 2.6: Gompertz–Verteilung*




a) $\gamma < 0$, b) $\gamma > 0$

*Abbildung 2.7: Log−logistische Verteilung*




a) $\gamma < 1$, b) $\gamma = 1$, $\gamma > 1$

fallendes oder steigendes Risiko mit allen drei Funktionstypen modelliert werden kann. Die log−logistische Rate kann zusätzlich für $\gamma > 1$ einen Prozeß beschreiben, in dem zunächst das Ereignisrisiko steigt und dann wieder fällt (z.B. Scheidungsrisiko, vgl. DIEKMANN/MITTER 1984b).

## *2.3 Verteilungsmodelle für Wartezeiten − Ein Überblick*

Der eine oder andere Leser wird sich vielleicht fragen, warum die zuvor genannten Raten derart komplexe Funktionen der Zeit sind. Kann man eine monoton wachsende Rate nicht auch einfacher parametrisieren als beispielsweise die log−logistische Rate? Die Erklärung liegt darin, daß die theoretische Statistik nicht den im vorherigen Abschnitt beschriebenen Weg (von der Rate zur Dichte) gegangen ist, sondern zunächst ein Verteilungsmodell im engeren Sinne spezifiziert hat − also die Dichte − und dann − wenn überhaupt − die Rate abgeleitet hat. Die einschlägigen Lehrbücher zur Verlaufsdatenanalyse beginnen daher in der Regel mit einer Darstellung unterschiedlicher Dichtefunktionen mit wundersam klingenden Namen und komplizierten Funktionsgleichungen, die allesamt der Modellierung von Wartezeitverteilungen dienen sollen. Da dieses mehr der Abschreckung als dem Verständnis dient (zumal in einer Einführung), habe ich alle Gleichungen in den Anhang B verbannt, den der interessierte Leser von Fall zu Fall zu Rate ziehen sollte. Ich möchte mich an dieser Stelle weniger auf eine enzyklopädische Aufzählung konzentrieren, sondern mehr auf die Frage

nach dem "Warum" bestimmter Vorgehensweisen. Allerdings geht auch dieses nicht ohne Mathematik.

Das Hauptproblem bei der Formulierung von Verteilungsmodellen für Wartezeiten läßt sich wie folgt zusammenfassen: Bei den meisten sozialen Prozessen häufen sich Ereignisse entweder zu Beginn oder am Ende des Prozesses. Das Resultat ist eine empirische Verteilung, die in der Regel schief ist. In den wenigsten Fällen tritt eine einigermaßen symmetrische Verteilung auf, bei der ober- und unterhalb des Mittelwertes in etwa gleich viel Fälle auftreten. Außerdem treten immer nur positive Werte auf, denn negative Zeiten kann es prinzipiell nicht geben. In dieser Situation liefert auch eine andere Skalierung z.B. durch Standardisierung (Abweichungen vom Mittelwert in Einheiten der Standardabweichung) wegen der mangelnden Symmetrie keine befriedigenden Ergebnisse. Das allseits bekannte Modell der *Normalverteilung* ist daher kaum geeignet, die Verteilung der ersten Tätigkeiten in Abbildung 2.1 zu beschreiben. Das gilt auch für alle anderen theoretischen Verteilungen, die wie die Normalverteilung die Form einer (nicht immer symmetrischen) Käseglocke haben: z.B. *Extremwert- oder logistische Verteilung*. Ihre Dichtefunktionen f(t) lassen sich durch zwei Parameter beschreiben, von denen der eine die *zentrale Lage* $\mu$ (location) und der andere die *Dispersion* $\sigma$ (scale) der Verteilung beschreibt. Wenn man bei dem Vergleich mit der Käseglocke bleiben will, dann entsprechen sie dem Griff und dem Durchmesser der Glocke.

Natürlich kann man dieses Problem ignorieren und trotzdem die Daten mit Hilfe einer Normal-, Extremwert- oder logistischen Verteilung beschreiben. Wenn aber der Mittelwert der zu beschreibenden Verteilung nicht groß genug ist, dann wird das jeweilige Verteilungsmodell negative Wartezeiten vorhersagen, die es eigentlich nicht geben darf[1]. Ganz abgesehen davon, wird die Verteilungsform nicht auf die Daten passen. In dieser Situation gibt es im Prinzip zwei Lösungen:

1. Man ändert die Verteilung, indem man die Verteilung einfach am Lageparameter $\mu$ abschneidet. Man verwendet sozusagen nur die rechte Hälfte der Verteilung.
2. Man transformiert die Zufallsvariable, so daß der ursprünglich beschränkte Wertebereich auf alle (positiven und negativen) reellen Zah-

1) Bei Annahme einer Normalverteilung müßte der Mittelwert 1,96-mal so groß sein wie die Standardabweichung, damit nur in 2,5% der Fälle ein negativer Wert auftreten kann.

len ausgeweitet wird. Eine solche Transformation ist z.B. der natürliche Logarithmus ln t. Wartezeiten zwischen 0 und 1 ergeben negative Werte (allerdings muß $t>0$ sein). Wartezeiten von 1 werden zu 0 und Wartezeiten größer als 1 ergeben positive Werte. Außerdem werden die Abstände im Bereich [0,1] vergrößert und im Bereich $t>1$ verkleinert. Wenn die Verteilung der ursprünglichen Wartezeiten bekannt ist, muß man dann nur die Frage klären, welche Verteilung die transformierte Variable haben wird. Diese Frage läßt sich auch umkehren: Welche Verteilung muß die ursprüngliche Wartezeit haben, damit auf die transformierte Variable eine der o.g. Verteilungen angewandt werden kann, die durch einen Lage– und einen Streuungsparameter gekennzeichnet sind?

Abschnitt 2.3.1 stellt zunächst die drei "Käseglocken"–Verteilungen vor und zeigt, wie man in diesen Fällen Überlebensfunktion und Rate berechnen würde, auch wenn dieses aus den genannten Gründen nicht besonders sinnvoll ist. In den beiden folgenden Abschnitten 2.3.2 und 2.3.3 möchte ich die beiden Lösungen (Transformation der Verteilung bzw. der Zufallsvariablen) kurz an Hand eines Beispiels erläutern, bevor ich in Abschnitt 2.3.4 einen allgemeinen Überblick über die gebräuchlichsten Verteilungsmodelle für Wartezeiten gebe. Es zeigt sich nämlich, daß viele der so fremd klingenden Verteilungen, wie z.B. die log–logistische oder die lognormale, nichts anderes als Transformationen bekannter Verteilungen sind, die durch eine der beiden genannten Strategien zustandegekommen sind.

### *2.3.1 Rate und Überlebensfunktion für Extremwert–, Normal– und logistische Verteilung*

Bei der *Extremwertverteilung* handelt es sich um eine theoretische Verteilung, die verwendet wird, um die Extremwerte einer Zufallsvariablen zu beschreiben (z.B. niedrigste Temperatur im Winter, Niederschlagsmenge in Dürreperioden etc., vgl. GUMPEL 1958). Die *Normalverteilung* ist aus Einführungen zur Statistik hinreichend bekannt. Sie gilt als Modell für Zufallsvariablen (z.B. Meßfehler), die sich aus vielen voneinander unabhängigen Größen zusammensetzen, von denen keine einen dominierenden Einfluß hat. Die *logistische Verteilung* wird häufig als Näherung der Normalverteilung verwendet, weil insbesondere ihre Verteilungsfunktion leichter zu berechnen ist.

Die Dichten aller drei Verteilungen sind in Anhang B zusammengefaßt. Sie sind für Zufallsvariablen definiert, deren Wertebereich von minus bis plus unendlich reicht. Zur besseren Unterscheidung bezeichne ich diese Zufallsvariablen mit Y. Für die Berechnung der Überlebenswahrscheinlichkeit aus der Dichte ist daher von $-\infty$ und nicht von 0 wie in (2.7b) zu integrieren:

$$S(y) = 1 - P(Y \leq y) = 1 - \int_{-\infty}^{y} f(u)\,du \tag{2.18}$$

Wenn man z.B. die Dichtefunktion der Extremwertverteilung einsetzt, ergibt sich die Formel der Überlebensfunktion wie folgt:

$$\begin{aligned} &S(y|\text{Extremwert}) \\ &= 1 - \int_{-\infty}^{y} \left\{ \frac{1}{\sigma} \exp\left( \frac{u-\mu}{\sigma} - \exp\frac{u-\mu}{\sigma} \right) \right\} du \\ &= \exp\left( -\exp\frac{y-\mu}{\sigma} \right) \end{aligned} \tag{2.19}$$

Diese Überlebenswahrscheinlichkeit hat bei minus unendlich den Wert 1. Die Rate erhält man schließlich unter Verwendung von (2.11b):

$$\begin{aligned} r(y|\text{Extremwert}) = \frac{f(y)}{S(y)} &= \frac{\frac{1}{\sigma} \exp\left( \frac{y-\mu}{\sigma} - \exp\frac{y-\mu}{\sigma} \right)}{\exp\left( -\exp\frac{y-\mu}{\sigma} \right)} \\ &= \frac{1}{\sigma} \exp\frac{y-\mu}{\sigma} \end{aligned} \tag{2.20}$$

S(y) und r(y) ergeben sich analog für Normal- und logistische Verteilung. An der Berechnung der drei charakteristischen Funktionen hat sich also prinzipiell nichts geändert, außer daß man den anderen Wertebereich der Zufallsvariablen Y berücksichtigen muß. Alle drei Funktionen sind in den folgenden Abbildungen dargestellt. Es ergibt sich für jedes Verteilungsmodell eine monoton steigende Rate.

*Abbildung 2.8: Standard – Normalverteilung*

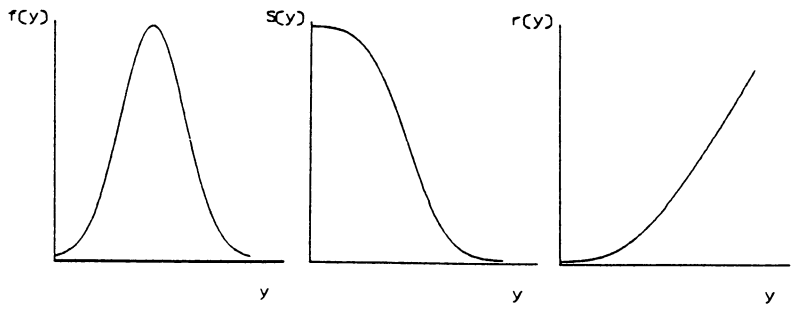



*Abbildung 2.9: Standard – Extremwertverteilung*

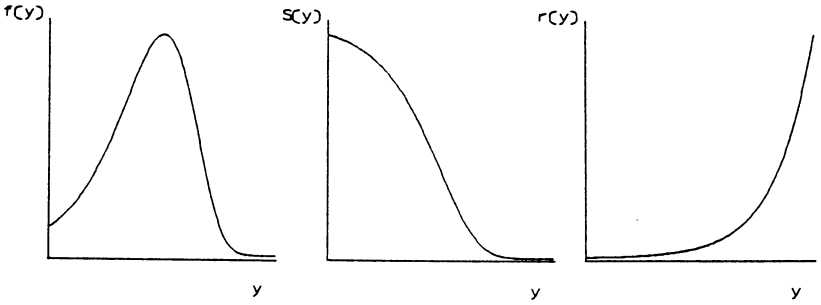
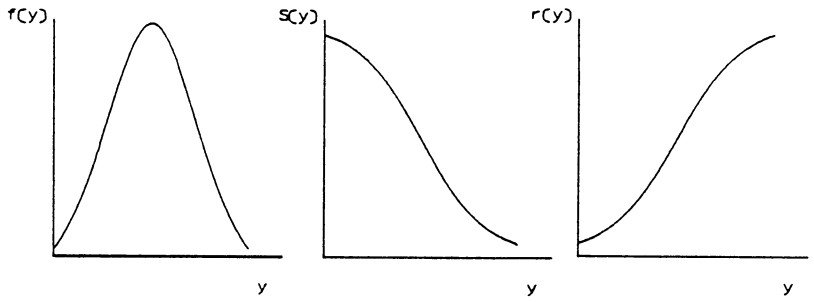



*Abbildung 2.10: Logistische Verteilung*



### 2.3.2 *Abgeschnittene Verteilungen*

Aus den genannten Gründen machen diese theoretischen Verteilungen jedoch für Wartezeiten keinen Sinn. Möchte man sie dennoch zur Beschreibung der positiven Zufallsvariablen T verwenden, dann muß man den linken Teil der Verteilung ignorieren. In diesem Fall berechnet man die Überlebenswahrscheinlichkeit wie gehabt nach (2.7b). Da jedoch der linke Teil der Verteilung abgeschnitten wird, ergibt das Integral der Dichte nicht mehr 1. Damit es sich weiter um eine Überlebenswahrscheinlichkeit handelt, muß die Verteilung auf die neuen Intervallgrenzen $[0,\infty)$ normiert werden (vgl. ELANDT–JOHNSON/JOHNSON 1980: 53 ff.). Für die Extremwertverteilung ergibt sich z.B.:

(2.21)

$$S(t) = 1 - \frac{\int_0^t f(u|\text{Extremwert})\,du}{\int_0^\infty f(u|\text{Extremwert})\,du} = \exp\left[\exp\frac{-\mu}{\sigma} - \exp\frac{t-\mu}{\sigma}\right]$$

Diese Überlebenswahrscheinlichkeit hat im Gegensatz zu (2.19) bei $t=0$ den Wert eins. Dabei handelt es sich um nichts anderes als die Überlebenswahrscheinlichkeit der *Gompertz–Verteilung*, wie man durch folgende Reparametrisierung sieht:

(2.22)
$$S(t|\text{Gompertz}) = \exp\left[\frac{\lambda}{\gamma}[1 - \exp(\gamma t)]\right]$$

$$\text{mit } \lambda = \sigma^{-1}\exp\frac{-\mu}{\sigma} \quad \text{und } \gamma = \sigma^{-1}$$

GOMPERTZ (1825) hat diese Verteilung zwar etwas anders abgeleitet, im Prinzip ist sie jedoch nichts anderes als eine abgeschnittene Extremwertverteilung. DAVIS (1952) sowie BARLOW und PROSCHAN (1965) verwenden diese Strategie, um Wartezeiten mit Hilfe einer (bei Null abgeschnittenen) *Standardnormalverteilung* zu analysieren.

### 2.3.3 *Transformation der Wartezeit*

Die Ausgangsfrage für diese zweite Alternative lautet: Welche Verteilung hat die ln−Transformation einer positiven Zufallsvariablen, deren eigene Verteilung bekannt ist? Angenommen man geht von *exponentiell verteilten Wartezeiten* aus. Deren Überlebensfunktion lautet bekanntermaßen:

$$S(t|\text{Exponential}) = \exp(-\lambda t)$$

Durch Einsetzen von $\exp(\ln t)=t$ wird diese Formel nicht verändert und nach einigen Umformungen erhält man:

$$\begin{aligned} S(t) &= \exp[-\lambda\exp(\ln t)] = \exp[-\exp(\ln\lambda)\exp(\ln t)] \\ &= \exp[-\exp(\ln\lambda + \ln t)] = \exp\left\{-\exp\left[\frac{\ln t - (-\ln\lambda)}{1}\right]\right\} \end{aligned}$$

Durch die Reparametrisierung $\mu=-\ln\lambda$ und $\sigma=1$ erhält man eine Formel, die der Überlebenswahrscheinlichkeit der Extremwertverteilung (2.19) entspricht:

$$S(y|\text{Extremwert}) = \exp\left[-\exp\frac{y-\mu}{\sigma}\right] \quad \text{mit } y = \ln t$$

Eine Verallgemeinerung der Exponentialverteilung ist die sogenannte *Weibull−Verteilung*. Sie wurde im wesentlichen entwickelt, um zeitabhängige Prozesse auf flexible Art und Weise zu modellieren. Ihre Überlebenswahrscheinlichkeit lautet:

$$S(t|\text{Weibull}) = \exp[-(\lambda t)^{\gamma}] \tag{2.23}$$

Sie enthält als Spezialfall ($\gamma=1$) die Exponentialverteilung. Auch hier kann man mit den gleichen Umformungen wie vorher zeigen, daß das Endergebnis einer ln−Transformation eine Extremwertverteilung ist:

$$S(t) = \exp\{-\exp[\ln(\lambda t)^{\gamma}]\} = \exp\{-\exp[\gamma(\ln\lambda + \ln t)]\}$$

Mit der Reparametrisierung $\mu = -\ln\lambda$, $\sigma = \gamma^{-1}$ und $y = \ln t$ ergibt sich:

$$S(y|\text{Extremwert}) = \exp\left[-\exp\frac{y-\mu}{\sigma}\right]$$

Allgemein kann man also sagen: Sind Wartezeiten Weibull (Spezialfall: exponentiell) verteilt, dann ist der natürliche Logarithmus der Wartezeiten extremwertverteilt. Aus diesem Grund bezeichnet man die Extremwertverteilung auch manchmal als *Log – Weibull – Verteilung.*

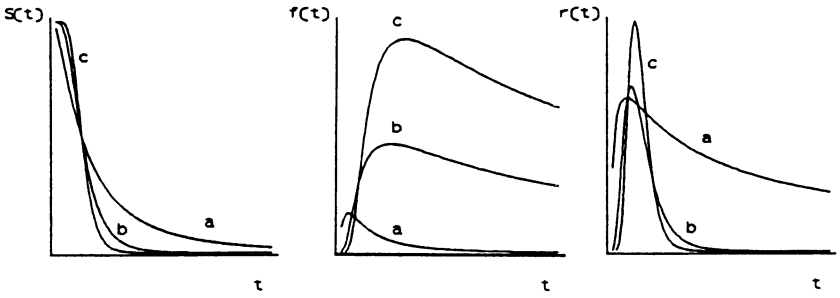
Die Fragestellung dieses Abschnitts läßt sich auch umkehren, indem man untersucht, welche Verteilung Wartezeiten haben müssen, damit man ihren natürlichen Logarithmus mit einer theoretischen Verteilung analysieren kann, deren Eigenschaften gut erforscht sind (z.B. die Normalverteilung oder die logistische Verteilung). In diesem Fall muß man auch die obige Ableitung umkehren und es ergeben sich die *log – normale und die log – logistische Verteilung* mit Überlebensfunktion:[1]

(2.24) $$S(t|\log-\text{normal}) = 1 - \Phi[\gamma \ln(\lambda t)]$$

(2.25) $$S(t|\log-\text{logistisch}) = (1 + (\lambda t)^{\gamma})^{-1}$$

Die log – logistische Verteilung habe ich bereits vorgestellt (vgl. Abbildung 2.7), zur Log – Normalverteilung vgl. die folgende Abbildung.

*Abbildung 2.11: Log – Normalverteilung*




a) $\gamma = 1$, b) $\gamma = 2$, c) $\gamma = 3$

1) $\Phi[.]$ ist das unvollständige Integral der Standard – Normalverteilung.

Sind also Wartezeiten log – normal oder log – logistisch verteilt, dann ist ihr natürlicher Logarithmus normal oder logistisch verteilt. Die Namensgebung mit der Vorsilbe log ist im Gegensatz zu dem vorherigen Beispiel irreführend: Dort ist die Wartezeit Weibull – und ihr natürlicher Logarithmus Log – Weibull – verteilt (= extremwertverteilt). Hier ist es genau umgekehrt.

### *2.3.4 Überblick über verschiedene Verteilungsmodelle*

Nach dieser "Vorrede" möchte ich die gebräuchlichsten Verteilungen kurz zusammenfassen. Ihre wesentlichen Eigenschaften sind in Form einer Synopse in Anhang B zusammengestellt. Es gibt also Verteilungen, die

a) direkt für Wartezeiten verwendet werden können, weil sie Zufallsvariablen beschreiben, die nur positive Werte annehmen können: Exponential –, Weibull –, Lognormal –, Log – logistische –, Gamma – Verteilung.

b) nur für logarithmische Transformationen der Wartezeit sinnvoll verwendet werden können, weil sie Zufallsvariablen beschreiben, die in ihrem Wertebereich unbeschränkt sind: Extremwert – Verteilung. In diesem Fall muß jedoch die ursprüngliche Wartezeit bestimmten Verteilungsannahmen genügen und Weibull – oder exponentiell verteilt sein.

c) deshalb direkt für Wartezeiten verwendet werden können, weil ein Teil der Verteilung ignoriert wird: Gompertz – Verteilung (abgeschnittene Extremwert – Verteilung).

In den Fällen a und c bezeichne ich die Zufallsvariable direkt mit T, bei b hingegen mit Y.

Ein Vergleich des Verlaufs der Dichtefunktionen zeigt auffällige Unterschiede zwischen den Verteilungen des Typs a und b. Wie schon erwähnt, haben die Verteilungen des Typs b die Form einer Glocke und können durch einen Lageparameter $\mu$ und einen Streuungsparameter $\sigma$ beschrieben werden. Wie das Beispiel der Extremwert – Verteilung zeigt, müssen sie nicht unbedingt symmetrisch sein. Ihre Verteilungsform ist jedoch im Gegensatz zum Typ a vergleichsweise einfach.

Bei Typ a finden sich dagegen die unterschiedlichsten Verteilungsformen. Auch hier hängt die Funktion f(t) von zwei Parametern ab, die jedoch

eine grundsätzlich andere Bedeutung haben. Der Parameter $\gamma$ bestimmt die Form (shape) der Verteilung und wird daher als *Formparameter* bezeichnet. Der zweite Parameter $\lambda$ beeinflußt nur die Skala der Zeit-Achse (Stauchung und Streckung der Achse in horizontaler Richtung) und wird daher als *Skalierungsparameter* bezeichnet.

Da alle Verteilungen vollständig bekannt sind, kann man bei gegebenen Parametern verschiedene Statistiken zur Beschreibung der Verteilung berechnen. Der Median der Verteilung ergibt sich z.B., indem man die Überlebenswahrscheinlichkeit gleich 0,5 setzt und nach t bzw. y auflöst. Die Formeln für Mittelwert und Varianz der jeweiligen Verteilung habe ich, soweit sie mir bekannt waren, ebenfalls in Anhang B aufgeführt. Dort finden sich auch Angaben über Vor- und Nachteile sowie Erweiterungen und Spezialfälle der einzelnen Verteilungen. Soweit es möglich war, wurde das stochastische Modell erwähnt, das Grundlage der einzelnen Verteilung ist.

Im Zusammenhang mit der Analyse von Verlaufsdaten ist insbesondere der Verlauf der *Rate* interessant, der sich mit den einzelnen Verteilungen ergibt. Wie die Abbildungen 2.4 bis 2.11 zeigen, sind dabei zunehmende, abnehmende und wechselnde Verläufe möglich. Für die Analyse eines zeitabhängigen Prozesses bieten sich also viele Alternativen.

Abschließend sei noch auf ein paar Literaturangaben hingewiesen, in denen sich weitere Angaben über die hier erwähnten Verteilungsmodelle finden. Fast jedes Standardwerk zur Survival Analysis enthält eine Übersicht über die gebräuchlichsten Modelle, beschränkt sich jedoch häufig auf die Verteilungen, die dann im Laufe der Abhandlung auch angewendet werden (vgl. z.B. GROSS/CLARK 1975, KALBFLEISCH/PRENTICE 1980, LAWLESS 1982). NELSON (1982) behandelt die verschiedenen Verteilungsmodelle für Wartezeiten m.E. nach am ausführlichsten, da sich sein Buch weniger mit Regressionsmodellen als mit der Anpassung von Verteilungen für empirische Daten beschäftigt. Eine umfassende Diskussion unter statistisch-theoretischen Gesichtspunkten findet sich bei COX/OAKES (1984). Einige Informationen über die stochastischen Modelle, die den verschiedenen theoretischen Verteilungen zugrundeliegen, erfährt man bei MANN et al. (1974). Ein Nachschlagewerk für die wichtigsten theoretischen Verteilungen der Statistik sind schließlich die drei Bände von JOHNSON und KOTZ (1970).

## 2.4 *Komplexere Veränderungsprozesse*

Bisher habe ich mich lediglich mit singulären Ereignissen beschäftigt. In diesem Fall kann sich die statistische Analyse auf eine Modellierung der Zeitdauer bis zum Eintritt des Ereignisses, also auf die Wartezeitverteilung beschränken. Viele in der Realität zu beobachtende Verlaufsprozesse sind jedoch komplexer: Häufig kann ein Ausgangszustand auf verschiedene Art und Weise beendet werden (multiple Ereignisse) und Zustandsänderungen können mehrmals im Zeitablauf auftreten (wiederholbare Ereignisse). In diesem Abschnitt soll die Modellierung dieser komplexeren Verlaufsprozesse soweit besprochen werden, wie es für die in Kapitel 4 und 5 vorgenommenen Auswertungen notwendig ist. Eine umfassendere Darstellung findet sich bei LANCASTER (1990: Kap. 5) und TUMA/HANNAN (1984: Kap. 3–4).

### 2.4.1 *Multiple Ereignisse*

Wenn ein Ausgangszustand j auf unterschiedliche Art und Weise beendet werden kann, dann ist neben der Zufallsvariablen $T_j$ noch die Zufallsvariable $Z_j(t)$ (Zielzustand) zu betrachten. Wenn insgesamt K verschiedene Ereignisse möglich sind, dann ist $Z_j(t)=k$ eine diskrete Zufallsvariable mit $k=1,\ldots,K$ Ausprägungen. Der Index j dient lediglich dazu, die Notation für eine mögliche Erweiterung auf verschiedene Ausgangszustände $j=1,\ldots,J$ offen zu halten. Innerhalb dieses Abschnitts wird der Ausgangszustand jedoch als gegeben vorausgesetzt.[1] Mit dieser erweiterten Notation lassen sich die bekannten charakteristischen Funktionen wie folgt schreiben:

(2.26a) $S_j(t) = P(T_j \geq t)$ Überlebensfunktion

(2.26b) $F_j(t) = 1 - S_j(t)$ Verteilungsfunktion

(2.26c) $f_j(t) = -dS_j/dt$ Dichte

1) Dies ist auch die Annahme aller empirischen Analysen in Kapitel 4 und 5. Praktisch kann man sich das so vorstellen, daß bei Auftreten mehrerer Ausgangszustände die Daten disaggregiert und getrennt für jeden Zustand $j=1,\ldots,J$ ausgewertet werden. Falls alle Untersuchungseinheiten im gleichen Ausgangszustand beginnen, kann dieser Index aber auch fortgelassen werden.

(2.26d) $r_j(t) = f_j(t)/S_j(t)$ Rate

Mit der Unterscheidung verschiedener Zielzustände $Z_j(t)$ kann man nun sowohl die bedingte Verteilung der Wartezeiten $T_j$ für jeden Zustand k als auch die gemeinsame Verteilung betrachten. Es ergeben sich entsprechende bedingte oder partielle Funktionen $S_{jk}(t)$ und $f_{jk}(t)$, jedoch haben sie nicht immer die in Abschnitt 2.2 besprochenen Eigenschaften einer Überlebens– oder Dichtefunktion. Ausgangspunkt dieser Überlegungen ist wiederum die *Rate*, die nunmehr nach verschiedenen Abgängen k aus dem Ausgangszustand j differenziert:

$$(2.27) \qquad r_{jk}(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T_j < t+\Delta t,\ Z_j(t)=k \mid T \geq t)}{\Delta t}$$

Sie wird daher als *Übergangsrate* oder *–intensität* bezeichnet, um sie auch begrifflich von der in Gleichung (2.9) definierten *Hazardrate* abzugrenzen. Wenn ich in beiden Fällen zwecks sprachlicher Vereinfachung den Terminus Rate verwende, sollte nicht vergessen werden, daß beide Funktionen unterschiedliche mathematische Eigenschaften haben.

Für unsere Überlegungen ist das Modell *konkurrierender Risiken* hilfreich. Das folgende Beispiel aus dem Bereich der Technik illustriert die wesentlichen Annahmen. In einem zweiten Schritt wäre dann zu fragen, ob dieses Modell auf sozialwissenschaftliche Fragestellungen übertragbar ist. Gegeben sei also ein Sicherungskasten, in dem drei Sicherungen *hintereinander* geschaltet sind. Der Strom wird offensichtlich dann unterbrochen, wenn eine der drei Sicherungen ausfällt. Die Wartezeit bis zu dieser Unterbrechung entspricht der *minimalen* Lebensdauer aller drei Sicherungen. Wenn man so will, "konkurrieren" die drei Sicherungen darum, welche zuerst ausfällt. Ganz im Gegensatz zu einem anderen Sicherungskasten, in dem diese Sicherungen *parallel* geschaltet sind. Dort wird der Strom erst dann unterbrochen, wenn die letzte Sicherung ausfällt. Der Zeitpunkt entspricht dann der *maximalen* Lebensdauer aller drei Sicherungen.

Diese Modellvorstellungen lassen sich auf epidemiologische Fragestellungen übertragen: Betrachten wir eine Stichprobe von 1000 Personen und drei mögliche Todesursachen, die im Zeitablauf mit konstanter Rate auftreten. Alle Personen beginnen also im Zustand $j=0$ (lebend) und können aus $K=3$ Gründen sterben. Wenn diese ursachenspezifischen Mortalitätsraten

die Werte $r_{01}(t)=\lambda_{01}=0{,}35$, $r_{02}(t)=\lambda_{02}=0{,}30$ und $r_{03}(t)=\lambda_{03}=0{,}20$ haben, dann kann man analog (2.13) eine "Überlebens"funktion definieren:

$$S^{*}_{jk}(t) = \exp\left[-\int r_{jk}(u)du\right], \tag{2.28}$$

die sich für die genannten zeitkonstanten Raten reduziert auf:

$$S^{*}_{jk}(t) = \exp\left(-\lambda_{jk}t\right)$$

Mit den zuvor genannten Werten für $\lambda_{jk}$ ist in Tabelle 2.2 berechnet, wieviel Personen $n_{jk}$ der Stichprobe zu Beginn eines Jahresintervalls die jeweilige Todesursache "überleben".

*Tabelle 2.2: Entwicklung der Risikomenge*

| | | Wechsel von j=0 nach | | | | Wechsel von j=0 nach | | |
|---|---|---|---|---|---|---|---|---|
| Intervall | Alle | k=1 | k=2 | k=3 | Alle | k=1 | k=2 | k=3 |
| t | $n_0$ | $n_{01}$ | $n_{02}$ | $n_{03}$ | $S_0(t)$ | $S_{01}(t)$ | $S_{02}(t)$ | $S_{03}(t)$ |
| 1 | 1000 | 1000 | 1000 | 1000 | 100% | 100% | 100% | 100% |
| 2 | 427 | 705 | 741 | 819 | 43% | 70% | 74% | 82% |
| 3 | 183 | 497 | 549 | 670 | 18% | 50% | 55% | 67% |
| 4 | 78 | 350 | 407 | 549 | 8% | 35% | 41% | 55% |
| 5 | 33 | 247 | 301 | 449 | 3% | 25% | 30% | 45% |
| 6 | 14 | 174 | 223 | 368 | 1% | 17% | 22% | 37% |
| 7 | 6 | 122 | 165 | 301 | 1% | 12% | 17% | 30% |
| 8 | 3 | 86 | 122 | 247 | 0% | 9% | 12% | 25% |
| 9 | 1 | 61 | 91 | 202 | 0% | 6% | 9% | 20% |
| 10 | 0 | 43 | 67 | 165 | 0% | 4% | 7% | 17% |
| | | | | | Basis | 1000 | 1000 | 1000 |

Wenn alle drei Todesursachen additiv wirken würden, dann wären innerhalb des ersten Jahres $(1000-705)+(1000-741)+(1000-819)=736$ Personen verstorben. Nach Ablauf eines weiteren Jahres wären schließlich auch die verbliebenen 264 Personen gestorben. Diese dramatische Reduktion der Fallzahl scheint wenig realistisch. In der Tat kann der Ausdruck (2.28) nur bedingt als Überlebenswahrscheinlichkeit interpretiert werden (deshalb die Anführungsstriche). Er wird daher auch als *Pseudo-Überlebensfunktion* bezeichnet. Er mißt die Überlebenswahrscheinlichkeit, wenn nur die Ursache k wirken würde. Die tatsächliche Überlebensfunktion

ergibt sich aus der Wahrscheinlichkeit, alle drei Todesursachen zu überleben. Macht man die vereinfachende Annahme, daß alle drei Todesursachen unabhängig voneinander wirken, dann entspricht diese Wahrscheinlichkeit dem Produkt der Einzelwahrscheinlichkeiten.

Gehen wir zunächst von der Gültigkeit dieser Annahme aus. Gegeben seien drei voneinander unabhängige, aber nicht beobachtbare Überlebenszeiten $U_1$, $U_2$ und $U_3$. Die beobachtete Wartezeit t bis zum Tod auf Grund einer der drei Ursachen entspricht der minimalen Überlebenszeit $t = \min(U_1, U_2, U_3)$. Bei Unabhängigkeit der drei Ereignisse gilt

$$P(T \geq t) = P(U_1 \geq t) \cdot P(U_2 \geq t) \cdot P(U_3 \geq t)$$

oder allgemein

$$(2.29) \quad S_j(t) = \prod_{k=1}^{K} \exp\left[-\int_0^t r_{jk}(u)du\right] = \exp\left[-\sum_{k=1}^{K}\int_0^t r_{jk}(u)du\right]$$
$$= \exp\left[-\int_0^t \sum_{k=1}^{K} r_{jk}(u)du\right]$$

Auf der Basis dieser Überlegungen wurden in Tabelle 2.2 zunächst die übergangsspezifischen Pseudo-Überlebenswahrscheinlichkeiten und dann die Überlebensfunktion der gesamten Stichprobe berechnet. Danach überleben am Ende des ersten Jahresintervalls 427 statt der zuvor berechneten 264 Personen – ein weitaus geringerer Rückgang der Risikomenge.

Eine weitere interessante Beziehung ergibt sich, wenn man die zeitkonstanten Raten $\lambda_{jk}$ in (2.29) einsetzt

$$S_j(t) = \exp\left[-\int_0^t \sum_{k=1}^{K} \lambda_{jk}\, du\right] = \exp\left[-\sum_{k=1}^{K} \lambda_{jk} t\right]$$

und einen Koeffizientenvergleich mit (2.17a) durchführt. Wenn man $\lambda$ in (2.17a) in der allgemeineren Notation als $\lambda_{j.}$ schreibt, gilt offenbar

$$\lambda_{j.} = \sum_{k=1}^{K} \lambda_{jk} = 0{,}35 + 0{,}30 + 0{,}20 = 0{,}85$$

Da zu einem Zeitpunkt t niemals zwei Ereignisse des gleichen Typs gleichzeitig auftreten können, gilt diese Beziehung ganz allgemein und nicht nur für zeitkonstante Raten:

$$r_{j.}(t) = \sum_{k=1}^{K} r_{jk}(t) \tag{2.30}$$

Die Hazardrate entspricht also der Summe der Übergangsraten über alle Zielzustände.[1] Mit $\lambda_{j.} = 0{,}85$ ergibt sich eine zweite Möglichkeit, die Überlebensfunktion $S_j(t) = \exp(-\lambda_{j.} t)$ zu berechnen. Sie führt zum gleichen Ergebnis wie die Multiplikation der übergangsspezifischen Pseudo-Überlebenswahrscheinlichkeiten, wie sich an Hand von Tabelle 2.2 leicht nachrechnen läßt.

Betrachten wir nun das Problem konkurrierender Risiken aus der Perspektive der Datenerhebung. Die übergangsspezifischen Risikomengen $n_{jk}$ aus Tabelle 2.2 können offensichtlich nicht beobachtet werden, da immer alle drei Risiken gleichzeitig wirken. Wie bereits festgestellt, kann man lediglich die minimale Überlebenszeit $t = \min(U_1, U_2, U_3)$ und damit nur die (globale) Überlebensfunktion $S_j(t)$ beobachten. Der Beobachter steht daher vor der Schwierigkeit, an Hand der erhobenen Veränderungen – getrennt nach Ursachen – inkl. deren Zeitpunkte auf die zugrundeliegenden konkurrierenden Risiken zu schließen. Diese Situation ist in Tabelle 2.3 dargestellt.

Demnach sterben im ersten Jahresintervall $1000 - 427 = 573$ Personen, wobei 236 Fälle auf die erste, 202 Fälle auf die zweite und 135 Fälle auf die dritte Todesursache zurückzuführen sind. In diesem Beispiel ist der Beobachter überdies in der glücklichen Lage, auch angeben zu können, wieviel Personen insgesamt an jeder Ursache sterben. In der Reihenfolge der drei Ursachen sind dies insgesamt 412, 353 bzw. 235 Personen. Die Wahrscheinlichkeit, an der ersten Ursache zu sterben, beträgt also $100 \cdot 412/1000 = 41$ Prozent. Normalerweise ist aber auf Grund eines begrenzten Untersuchungszeitraums, in dem ein Teil der Beobachtungen zensiert ist, diese Wahrscheinlichkeit nicht direkt beobachtbar.

---

1) Diese Eigenschaft wird durch den Punkt als zweiten Index bei $r_{j.}(t)$ verdeutlicht. Der Punkt wird in dieser Arbeit immer dann verwendet, wenn über einen Index summiert wurde.

*Tabelle 2.3: Multiple Ereignisse*

| | Wechsel von j=0 nach | | | | Wechsel von j=0 nach | | | |
|---|---|---|---|---|---|---|---|---|
| Intervall | Alle | k=1 | k=2 | k=3 | Alle | k=1 | k=2 | k=3 |
| t | $d_0$ | $d_{01}$ | $d_{02}$ | $d_{03}$ | $F_0(t)$ | $F^*_{01}(t)$ | $F^*_{02}(t)$ | $F^*_{03}(t)$ |
| 1 | 573 | 236 | 202 | 135 | 57% | 24% | 20% | 13% |
| 2 | 245 | 101 | 86 | 58 | 82% | 34% | 29% | 19% |
| 3 | 105 | 43 | 37 | 25 | 92% | 38% | 33% | 22% |
| 4 | 45 | 18 | 16 | 11 | 97% | 40% | 34% | 23% |
| 5 | 19 | 8 | 7 | 4 | 99% | 41% | 35% | 23% |
| 6 | 8 | 3 | 3 | 2 | 99% | 41% | 35% | 23% |
| 7 | 3 | 1 | 1 | 1 | 100% | 41% | 35% | 23% |
| 8 | 1 | 1 | 1 | 0 | 100% | 41% | 35% | 24% |
| 9 | 1 | 0 | 0 | 0 | 100% | 41% | 35% | 24% |
| 10 | | | | | | | | |
| Summe | 1000 | 412 | 353 | 235 | Basis | 1000 | 1000 | 1000 |

An Hand der Daten aus Tabelle 2.3 kann der Beobachter übergangsspezifische Verteilungsfunktionen berechnen, die wie folgt definiert sind:

$$F^*_{jk}(t) = P(T<t,\ Z_j(t)=k) \tag{2.31}$$

Ein Schätzer dieser Funktion ist der kumulative Anteil der Sterbefälle einer Ursachengruppe an allen Sterbefällen (vgl. Tabelle 2.3). Die (globale) Verteilungsfunktion $F_j(t)$ ergibt sich schließlich durch Summierung der übergangsspezifischen Funktionen

$$F_j(t) = \sum_{k=1}^{K} F^*_{jk}(t) \tag{2.32}$$

Im Gegensatz zur globalen haben jedoch die übergangsspezifischen Funktionen nicht die Eigenschaften einer Verteilungsfunktion, die eine monoton zunehmende Funktion mit Endwert $F(\infty)=1$ sein muß. Sei $\pi_{jk}$ die Wahrscheinlichkeit, irgendwann einmal an der Ursache k zu sterben, dann gilt für den Endwert der übergangsspezifischen Verteilungsfunktion

$$F^*_{jk}(\infty) = \pi_{jk}$$

Aus diesem Grund wird $F^*_{jk}(t)$ auch als *Sub-Verteilungsfunktion* bezeichnet. Die gleichen Überlegungen gelten im übrigen für die übergangsspezifischen Dichten $f^*_{jk}(t) = dF^*_{jk}(t)/dt$, die wie folgt definiert sind:

$$(2.33) \qquad f^*_{jk}(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T_j < t + \Delta t,\ Z_j(t) = k)}{\Delta t}$$

Hier ist das Integral über $f^*_{jk}(t)$ nicht 1, wie es üblicherweise für eine Dichtefunktion vorausgesetzt wird, sondern gleich $\pi_{jk}$.

Würde man die Wahrscheinlichkeit $\pi_{jk}$ kennen, dann könnte man $F^*_{jk}(t)$ und $f^*_{jk}(t)$ entsprechend korrigieren. $r_{jk}(t)dt$ ist nach den obigen Definitionen die bedingte Wahrscheinlichkeit, in einem Intervall $[t,t+dt)$ nach k zu wechseln, vorausgesetzt die Untersuchungseinheit hat bis t überlebt. Die unbedingte Wahrscheinlichkeit eines Wechsels nach k im Intervall $[t,t+dt)$ entspricht dann $S_j(t)r_{jk}(t)dt$. Die Wahrscheinlichkeit $\pi_{jk}$, überhaupt nach k zu wechseln, ergibt sich schließlich durch Integration über alle Zeitpunkte:[1]

$$(2.34) \qquad \pi_{jk} = \int_0^\infty S_j(u) r_{jk}(u) du$$

Bei zeitkonstanten Übergangsraten entspricht diese Wahrscheinlichkeit

$$\pi_{jk} = \int_0^\infty \exp(-\lambda_j u)\, \lambda_{jk} du = \lambda_{jk} / \lambda_j.$$

wie man leicht nachrechnen kann. Z.B. erhält man für die erste Ursache mit $\lambda_{01}/\lambda_0 = 0{,}35/0{,}85 = 0{,}41$ die gleiche Wahrscheinlichkeit wie zuvor mit den Häufigkeiten 412/1000.

Mit den $\pi_{jk}$ lassen sich nunmehr bedingte Funktionen für verschiedene Übergänge berechnen, die wieder die gewünschten Eigenschaften einer Dichte- bzw. Verteilungsfunktion haben:

$$(2.35) \qquad f_{jk}(t) = f^*_{jk}(t)/\pi_{jk}$$

$$(2.36) \qquad F_{jk}(t) = F^*_{jk}(t)/\pi_{jk} = P(T < t \mid Z_j(t) = k)$$

1) Mit der Beziehung (2.30) impliziert (2.34) übrigens $\sum_k \pi_{jk} = 1$.

Ein Schätzer dieser bedingten übergangsspezifischen Verteilungsfunktion (2.36) ist der kumulative Anteil der Sterbefälle einer Ursachengruppe an allen Sterbefällen der Ursachengruppe. Das Komplement dieser Funktion

$$(2.37) \qquad S_{jk}(t) = 1 - F_{jk}(t) = P(T \geq t_i | Z_j(t) = k)$$

entspricht dann einer Überlebensfunktion mit der gewünschten Eigenschaft $S_{jk}(\infty) = 0$. Beide Funktionen sind in der folgenden Tabelle 2.4 für die Beispieldaten berechnet.

*Tabelle 2.4: Verteilungs- und Überlebensfunktion*

| | | Wechsel von j=0 nach | | | | Wechsel von j=0 nach | | |
|---|---|---|---|---|---|---|---|---|
| Intervall | Alle | k=1 | k=2 | k=3 | Alle | k=1 | k=2 | k=3 |
| | $F_0(t)$ | $F_{01}(t)$ | $F_{02}(t)$ | $F_{03}(t)$ | $S_0(t)$ | $S_{01}(t)$ | $S_{02}(t)$ | $S_{03}(t)$ |
| 1 | 57% | 57% | 57% | 57% | 100% | 100% | 100% | 100% |
| 2 | 82% | 82% | 82% | 82% | 43% | 43% | 43% | 43% |
| 3 | 92% | 92% | 92% | 92% | 18% | 18% | 18% | 18% |
| 4 | 97% | 97% | 97% | 97% | 8% | 8% | 8% | 8% |
| 5 | 99% | 99% | 99% | 99% | 3% | 3% | 3% | 3% |
| 6 | 99% | 99% | 99% | 99% | 1% | 1% | 1% | 1% |
| 7 | 100% | 100% | 100% | 100% | 1% | 1% | 1% | 1% |
| 8 | 100% | 100% | 100% | 100% | 0% | 0% | 0% | 0% |
| 9 | 100% | 100% | 100% | 100% | 0% | 0% | 0% | 0% |
| 10 | | | | | 0% | 0% | 0% | 0% |
| | Basis | 412 | 353 | 235 | Basis | 412 | 353 | 235 |

In der Tabelle sind auch noch einmal globale Verteilungs- und Überlebensfunktion ausgewiesen. Wie man ebenfalls leicht nachvollziehen kann, lassen sie sich auch aus einer gewichteten Summe der bedingten übergangsspezifischen Funktionen berechnen:

$$(2.38) \qquad F_j(t) = \sum_{k=1}^{K} \pi_{jk} F_{jk}(t)$$

$$(2.39) \qquad S_j(t) = \sum_{k=1}^{K} \pi_{jk} S_{jk}(t)$$

Hierbei fungieren die Wahrscheinlichkeiten $\pi_{jk}$ als Gewichte. Eine solche

Verteilung, die sich additiv aus mehreren anderen zusammensetzt, bezeichnet man auch als *Mischverteilung*. Die insgesamt beobachtete Wartezeitverteilung setzt sich also aus verschiedenen bedingten Verteilungen zusammen (zur Abgrenzung häufig als "holding time distributions" bezeichnet).

Nach dieser fast unüberschaubaren Definition unterschiedlicher bedingter und unbedingter Funktionen faßt Tabelle 2.5 die wesentlichen Ereignisse und ihre Wahrscheinlichkeiten zusammen. Uns wird im folgenden vor allem die Übergangsrate interessieren. Sie ist nach Tabelle 2.5 (vgl. die Definitionen 2.27 und 2.33 sowie die Umformung 2.35):

$$(2.40) \qquad r_{jk}(t) = f^{*}_{jk}(t) \,/\, S_j(t) = \pi_{jk}\, f_{jk}(t) \,/\, S_j(t)$$

Man beachte, daß $r_{jk}(t)$ nicht gleich $f_{jk}(t)/S_{jk}(t)$ ist, was der Fall sein müßte, wenn $r_{jk}(t)$ eine (bedingte) Hazardrate wäre. Das bedingende Ereignis für $r_{jk}(t)$ ist das Überleben bis t und nicht Überleben bis t und Wechsel nach k.

*Tabelle 2.5: Zusammenfassung*

| Wahrscheinlichkeit | Definition | Ereignis | Bedingung |
|---|---|---|---|
| $r_{jk}(t)dt$ | (2.27) | Wechsel nach k in t | Überleben bis t |
| $r_{j.}(t)dt$ | (2.26) | Wechsel in t | Überleben bis t |
| $r_{jk}(t)/r_{j.}(t)$ | (2.42) | Wechsel nach k | Wechsel in t |
| $\pi_{jk}$ | (2.34) | Wechsel nach k | ... |
| $f_{jk}(t)dt$ | (2.35) | Wechsel in t | Wechsel nach k |
| $f^{*}_{jk}(t)dt$ | (2.33) | Wechsel nach k in t | ... |
| $f_{j}(t)dt$ | (2.26) | Wechsel in t | ... |
| $F_{jk}(t)$ | (2.36) | Wechsel vor t | Wechsel nach k |
| $F^{*}_{jk}(t)$ | (2.31) | Wechsel nach k vor t | ... |
| $F_{j}(t)$ | (2.26) | Wechsel vor t | ... |
| $S_{jk}(t)$ | (2.37) | Überleben bis t | Wechsel nach k |
| $S_{j}(t)$ | (2.26) | Überleben bis t | ... |
| $S^{*}_{jk}(t)$ | (2.28) | Überleben bis t | k einzige Ursache |

Für die empirischen Analysen mit multiplen Ereignissen werde ich auf die Raten $r_{j.}(t)$ bzw. $r_{jk}(t)$ und die davon ableitbaren Überlebensfunktionen $S_j(t)$ bzw. $S_{jk}(t)$ sowie die Sub-Dichte $f^{*}_{jk}(t)$ zurückgreifen. Wie aus der Zusammenfassung zu ersehen ist, mißt $f^{*}_{jk}(t)dt$ die Wahrscheinlichkeit eines

Wechsels nach k im Intervall [t,t+dt). Diese Wahrscheinlichkeit entspricht nach (2.40) dem Produkt $S_j(t)r_{jk}(t)dt$. $f^*_{jk}(t)$ kann daher unter Zuhilfenahme von (2.29) ebenfalls als Funktion der Übergangsraten betrachtet werden:

$$(2.41) \qquad f^*_{jk}(t) = r_{jk}(t) \exp\left[ -\int_0^t \sum_{k=1}^{K} r_{jk}(u)du \right]$$

Ähnlich wie bei singulären Ereignissen lassen sich also auch hier alle wesentlichen Funktionen auf Raten zurückführen.

Bevor ich mit der allgemeinen Diskussion fortfahre, seien zwei Probleme des Modells konkurrierender Risiken angemerkt. Wie bereits erwähnt, können die übergangsspezifischen Risikomengen nicht beobachtet werden. Dies erklärt, daß nur unter bestimmten vereinfachenden Annahmen auf die zugrundeliegenden Risiken geschlossen werden kann. Inbesondere ist es nicht möglich, allein auf Basis der (globalen) Überlebensfunktion $S_j(t)$ (vgl. Tabelle 2.2) zu entscheiden, ob die verschiedenen Risiken nun unabhängig voneinander wirken (wie in 2.29 unterstellt) oder nicht.[1] Da die Unabhängigkeitsannahme die statistischen Ableitungen wesentlich vereinfacht, wird die Gültigkeit dieser Annahme häufig unterstellt.[2]

Bei technischen Anwendungen mag dieses Vorgehen häufig gerechtfertigt sein: Ersetzt man beispielsweise eine der drei Sicherungen durch ein Modell, das nie ausfällt, dann dürfte die Lebensdauer der beiden anderen Sicherungen davon relativ unbeeinflußt sein. Anders verhält es sich jedoch schon bei epidemiologischen Untersuchungen: Schon sehr früh hat man sich dort gefragt, wie sich die Mortalität entwickeln würde, wenn es gelänge, bestimmte Risiken zu eliminieren.[3] Gerade aus der modernen Krebs-

---

1) Genauer gesagt, können ein Modell A, das von abhängigen Risiken ausgeht, und ein Modell B, das unabhängige Risiken unterstellt, die gleiche Randverteilung $S_j(t)$ haben, so daß ohne zusätzliche Annahmen nicht zwischen beiden unterschieden werden kann. Ein formaler Beweis dieses Identifikationsproblems würde den Rahmen dieser Arbeit sprengen, vgl. jedoch TSIATIS (1978). Auch die Kenntnis der ursachenspezifischen Ereignisse und deren Zeitpunkte löst dieses Problem nicht, zumal in allen realen Erhebungssituationen ein Teil der Beobachtungen kein Ereignis aufweisen wird.

2) Für die Auswertung von Verlaufsdaten hat diese Annahme u.a. den praktischen Vorteil, daß die einzelnen Übergangsraten unabhängig voneinander geschätzt werden können (vgl. Abschnitt 5.3.1).

3) DAVID und MOESCHBERGER (1978) erinnern in diesem Zusammenhang an D. Bernoulli's Überlegungen (1760) zur Elimination der Volksseuche Pocken.

forschung ist jedoch bekannt, wie sich verschiedene Mortalitätsrisiken gegenseitig verstärken oder vermindern können. KALBFLEISCH/PRENTICE (1980: 175ff.) berichten etwa über die immunologischen Abwehrreaktionen bei Leukämiepatienten, denen Knochenmark transplantiert wurde. Sie stellen fest, daß das Risiko, an den Folgen der Immunabwehr, d.h. an der Therapie selbst zu sterben, umgekehrt das Leukämierisiko vermindert.

Auch in den Sozialwissenschaften ist die Unabhängigkeit der Risiken häufig zweifelhaft. Man denke etwa an Arbeitslose: Sie finden entweder eine neue Arbeit oder scheiden aus dem Arbeitsmarkt aus. Wenn die Wiederbeschäftigungschancen sich verschlechtern (z.B. bei Frauen und älteren Arbeitnehmern in einem "enger" werdenden Arbeitsmarkt), dann ist anzunehmen, daß sich die Abgangsrate aus dem Arbeitsmarkt erhöht, das Risiko des "Ausscheidens" also nicht unabhängig vom "Risiko" der "Wiederbeschäftigung" ist. Auf Grund des genannten Identifikationsproblems muß jedoch immer wieder betont werden, daß unsere Möglichkeiten, die Abhängigkeit zu testen, sehr beschränkt sind.[1]

Eine zweite Anmerkung bezieht sich auf die Verteilung der Ereignisse nach Ursachen in Tabelle 2.3. Für das Beispiel mußten sie auf eine bestimmte Art und Weise berechnet werden, da aus der (globalen) Überlebensfunktion $S_j(t)$ noch nicht zu erkennen ist, wie sie sich auf die einzelnen Ursachen verteilen.[2] Wir verwenden stattdessen die bedingte Wahrscheinlichkeit $m_{jk}(t)$ eines Wechsels in den Zustand k, vorausgesetzt zum Zeitpunkt t findet ein Wechsel statt. Nach den Definitionen (2.26d) und (2.27) entspricht diese Wahrscheinlichkeit

$$m_{jk}(t) = r_{jk}(t) / r_{j.}(t) \tag{2.42}$$

Wenn nun, wie oben angenommen, alle drei Übergangsraten zeitkonstant sind, dann ist diese bedingte (zeitpunktspezifische) Wechselwahrscheinlichkeit ebenfalls konstant und für alle Zeitpunkte gleich. In diesem speziellen Fall entspricht sie damit der (globalen) Wahrscheinlichkeit $\pi_{jk}$, überhaupt in den Zustand k zu wechseln (vgl. 2.34).

---

1) Vgl. dazu KALBFLEISCH/PRENTICE (1980: 175ff.) und GALLER (1988) sowie zu den Konsequenzen abhängiger Risiken KLEIN (1988).

2) Auch die Pseudo-Überlebensfunktionen $S^*_{jk}(t)$ helfen hier nicht weiter, selbst wenn man sie beobachten könnte.

Dementsprechend ergibt sich in Tabelle 2.3 für jedes Zeitintervall der gleiche Anteil von Ereignissen für eine Ursache (z.B. immer 41% für Ursache 1). Anders ausgedrückt, die Wahrscheinlichkeit, daß eine Person in den Zustand k wechselt, hängt nicht davon ab, wann die Person wechselt, und die ursachenspezifischen Verteilungen sind alle identisch. Diese Eigenschaft gilt im übrigen für eine sehr viel allgemeinere Klasse von Übergangsraten, wenn diese zu jedem Zeitpunkt ein konstantes Vielfaches $m_{jk}$ der (globalen) Abgangsrate sind (vgl. LANCASTER 1990: 103f.):

$$(2.43) \qquad r_{jk}(t) = m_{jk} r_{j.}(t) \quad \Leftrightarrow \quad r_{jk}(t) / r_{j.}(t) = m_{jk}$$

Man bezeichnet sie als *Modelle mit proportionalen Übergangsraten.*[1]

### *2.4.2 Wiederholbare Ereignisse*

Wie in der Einführung dargestellt, ist das mehrfache Auftreten von Zustandsänderungen in den Sozialwissenschaften ein realistischer Anwendungsfall, der es erlaubt, eine Menge zusätzlicher Hypothesen über den Verlaufsprozeß zu testen. Ich habe sie unter dem Oberbegriff "Einflüsse der Vorgeschichte" zusammengefaßt. Gemeint sind damit Prozesse, in denen die Wahrscheinlichkeit bestimmter Zustandsänderungen von der Häufigkeit früherer Ereignisse oder der kumulierten Dauer vorhergehender Zustände abhängt. Erstmalig spielt damit nicht nur die Wartezeit t innerhalb des Zustands eine Rolle, sondern auch der Zeitpunkt s, seit dem dieser Zustand eingenommen wurde.

Eine besonders einfache Klasse von Modellen ergibt sich, wenn man lediglich ein bestimmtes Ereignis betrachtet, das sich im Zeitablauf wiederholt: z.B. die Anzahl $d_i$ der Tätigkeitswechsel, die bei einer Untersuchungseinheit i im Untersuchungszeitraum zu beobachten sind. In diesem Fall kann man zeigen (vgl. LAWLESS 1982: 494ff., LANCASTER 1990: 85ff.), daß die Häufigkeitsverteilung eines solchen wiederholbaren, singulären Ereignisses im Untersuchungszeitraum [0,t] einer Poissonverteilung mit Parameter $\mu = \lambda t$

1) Nicht zu verwechseln mit dem Modell proportionaler Risiken (vgl. Kapitel 5), in dem sich das Risiko eines bestimmten Übergangs von j nach k für zwei Personen mit unterschiedlichen Eigenschaften proportional verhält.

$$f(d_i) = \frac{(\lambda t)^{d_i} \exp(-\lambda t)}{d_i!} \tag{2.44}$$

entspricht, wenn Ereignisse mit konstanter Rate $r(t)=\lambda$ auftreten. Man spricht daher auch von einem *Poisson-Prozeß*. Der Erwartungswert einer Poissonverteilung entspricht bekanntlich ihrem Parameter $\mu$. Von daher ist zu erwarten, daß innerhalb des Untersuchungszeitraums [0,t] bei einer Untersuchungseinheit durchschnittlich

$$E(d_i) = \lambda t \tag{2.45}$$

Ereignisse auftreten. In einer Stichprobe vom Umfang N sind es dann insgesamt $D=N\lambda t$ Ereignisse.[1] Allgemeinere Poisson-Prozesse ergeben sich, wenn man zeitabhängige Raten unterstellt. Alle Poisson-Prozesse haben die Eigenschaft, daß die Zustandsdauer im Zustand Nr. (i+1) unabhängig von der Zustandsdauer im Zustand Nr. (i) ist. Sind Zustandsdauern voneinander unabhängig und gleich verteilt, dann spricht man auch von einem *Erneuerungsprozeß* (vgl. LAWLESS 1982: 496ff., LANCASTER 1990: 88ff.).

Betrachtet man schließlich mehrfache Zustandsänderungen zwischen verschiedenen Ausgangs- und Zielzuständen, dann ist im Gegensatz zum vorherigen Abschnitt nicht nur der Zielzustand $Z(t)=k$ bei einem Wechsel zum Zeitpunkt t, sondern auch der Ausgangszustand $Z(s)=j$ zu Beginn s des Zustands eine Zufallsvariable. Seien $q_j(s)$ die $j=1,\ldots,J$ Wahrscheinlichkeiten, sich zum Zeitpunkt s in einem der J Ausgangszustände zu befinden, und $r_{jk}(t)dt$ die entsprechenden Übergangswahrscheinlichkeiten, innerhalb des Intervalls [s,s+dt) in den Zustand k zu wechseln, dann läßt sich die Zustandsverteilung zum Zeitpunkt (s+dt) leicht fortschreiben, vorausgesetzt die Übergangswahrscheinlichkeiten sind nur von dem aktuell eingenommenen Zustand j und nicht von früheren Zuständen abhängig.

Dies ist die Standardannahme eines *Markov-Prozesses*. Die Grundüberlegung ist dabei in etwa folgende: Wenn man eine Momentaufnahme des Prozesses zum Zeitpunkt s macht, dann kann man die Veränderungen im nächsten Moment mit Hilfe der Übergangsraten $r_{jk}(t)$ prognostizieren, wenn die Vorgeschichte keine Rolle spielt. Eine Erweiterung dieses Mo-

1) Einige sozialwissenschaftliche Anwendungen für Modelle mit Ereignishäufigkeiten findet man bei COLEMAN (1981), KING (1989) und ANDRESS (1989).

dells ergibt sich, wenn sich die Übergangsraten mit der aktuellen Zustandsdauer t verändern. In diesem Fall spricht man von einem *Semi–Markov–Prozeß*. Falls schließlich Effekte der Vorgeschichte eine Rolle spielen, dann muß dies in der funktionalen Abhängigkeit der Übergangsraten berücksichtigt werden.

Man sieht, auch bei der Analyse wiederholbarer Ereignisse spielen Raten eine zentrale Rolle. Da wir uns im empirischen Teil dieser Arbeit im wesentlichen auf die Analyse der ersten Ereignisse innerhalb eines Verlaufs beschränken werden, möchte ich mich mit diesen allgemeinen Bemerkungen begnügen. Interessierte Leser finden weitere Erläuterungen insbesondere in der sozialwissenschaftlich orientierten Grundlagenliteratur (LANCASTER 1990 und TUMA/HANNAN 1984). Dort wird vor allem gezeigt, wie man mit Hilfe von Übergangsraten verschiedene Aspekte von Markov– und Semi–Markov–Prozessen vorhersagen kann – also z.B. die Zustandsverteilung zu einem bestimmten Zeitpunkt, die Anzahl bestimmter Zustandsänderungen in einem beliebigen Zeitintervall, die Übergangswahrscheinlichkeit zwischen zwei Zuständen j und k usw..

## *2.5 Einige Komplikationen*

Bei allen vorhergehenden Beispielen bin ich davon ausgegangen, daß der analysierte soziale Prozeß homogen ist und vollständig beobachtet werden konnte. Es ist jedoch anzunehmen, daß die Rate nicht nur im Zeitablauf variiert, sondern auch je nach Merkmalen der Untersuchungseinheiten größer oder kleiner ist. Diese vereinfachende Annahme diente eigentlich nur dazu, die Beispiele überschaubar zu halten. Für alle praktisch relevanten Anwendungen ist es jedoch notwendig, sowohl Veränderungen der Raten im Zeitablauf als auch Unterschiede zwischen den Untersuchungseinheiten zu berücksichtigen. Dabei ist zu berücksichtigen, daß in allen praktischen Anwendungsfällen nur ein Teil der Beobachtungen vollständig, d.h. bis zum Eintritt eines Ereignisses beobachtet werden kann.

Zeitabhängige Raten habe ich schon in Abschnitt 2.2 erwähnt. Wie ich jedoch in Abschnitt 2.5.1 zeigen möchte, hat die Existenz individueller Unterschiede zwischen den Untersuchungseinheiten zur Folge, daß zeitkonstante Prozesse als zeitabhängig erscheinen, wenn man diese heterogenen Subpopulationen nicht kontrolliert. Jedes realistische Erklärungsmodell

eines sozialen Prozesses sollte daher zunächst die Heterogenität des Untersuchungsmaterials berücksichtigen. In Abschnitt 2.5.2 werde ich schließlich die Problematik zensierter Beobachtungen näher beleuchten.

### *2.5.1 Heterogene Subpopulationen*

Eine Zeichnung der Rate gibt erste Aufschlüsse darüber, ob es sich um einen zeitkonstanten oder einen zeitabhängigen Prozeß handelt (vgl. Abbildung 2.3). Kommen wir daher noch einmal auf das Beispiel aus Abschnitt 2.2 zurück, in dem das Auftreten von Tätigkeitswechseln das Thema war (singuläre, nicht-wiederholbare Ereignisse). Die Daten eines ähnlichen Beispiels sind in aggregierter Form in Tabelle 2.6 dargestellt. Man erkennt, daß die (zeitdiskrete) Rate für diese Stichprobe im Untersuchungszeitraum deutlich abnimmt. Innerhalb von 5 Jahren ist sie von 0,244 um fast die Hälfte auf 0,131 zurückgegangen. Man könnte also sagen, die Karrieremobilität nimmt mit zunehmender Berufserfahrung ab. Bei näherem Hinsehen erweist sich diese Schlußfolgerung jedoch als falsch. Disaggregiert man nämlich die Untersuchungsgruppe nach Männern und Frauen, dann zeigt sich, daß Frauen zwar eine sehr viel höhere Mobilitätsrate als Männer aufweisen, in beiden Subgruppen das Ereignisrisiko aber konstant ist. Unterschiede der Karrieremobilität zwischen den Personen haben also zur Folge, daß auf der Ebene der gesamten Untersuchungsgruppe der Eindruck entsteht, als würde die Karrieremobilität generell abnehmen.

Allgemein gilt, daß mangelnde Kontrolle heterogener Subpopulationen scheinbare Zeitabhängigkeiten auf der Ebene der gesamten Stichprobe erzeugt und das Ereignisrisiko (scheinbar) im Zeitablauf abnimmt. Angenommen, individuelle Wartezeiten sind exponentiell verteilt, wobei das Ereignisrisiko $\lambda_i$ für jede Untersuchungseinheit i verschieden sein soll ($\lambda_i > 0$). Die Dichteverteilung der $\lambda_i$ sei $g(\lambda)$, deren Form unbekannt sei. Dichte und Verteilungsfunktion der Wartezeiten für die gesamte Untersuchungsgruppe ergeben sich dann durch Integration der individuellen Dichten:

$$f(t) = \int_0^\infty \lambda \exp(-\lambda t)\, g(\lambda)\, d\lambda \tag{2.46a}$$

(2.46b) $$S(t) = \int_0^\infty \exp(-\lambda t)\, g(\lambda)\, d\lambda$$

Aus den beiden Funktionen f(t) bzw. S(t) läßt sich nach (2.11) die Rate r(t) berechnen:

(2.47) $$r(t) = \frac{f(t)}{S(t)} = \frac{\int_0^\infty \lambda \exp(-\lambda t)\, g(\lambda)\, d\lambda}{\int_0^\infty \exp(-\lambda t)\, g(\lambda)\, d\lambda}$$

Für diesen Ausdruck läßt sich zeigen (vgl. z.B. FLINN/HECKMAN 1982:43), daß die erste Ableitung dr(t)/dt der Rate nach der Zeit (umgangssprachlich: die Veränderung der Rate) immer negativ ist.

*Tabelle 2.6: Erste Tätigkeitswechsel nach Jahren und Geschlecht (simulierte Daten)*

| Jahr | Alle | | | Männer | | | Frauen | | |
|---|---|---|---|---|---|---|---|---|---|
| | $n_k$ | $d_k$ | $\hat{q}_k$ | $n_k$ | $d_k$ | $\hat{q}_k$ | $n_k$ | $d_k$ | $\hat{q}_k$ |
| 1 | 2000 | 488 | 0,244 | 1000 | 95 | 0,095 | 1000 | 393 | 0,393 |
| 2 | 1512 | 325 | 0,215 | 905 | 86 | 0,095 | 607 | 239 | 0,394 |
| 3 | 1187 | 223 | 0,188 | 819 | 78 | 0,095 | 968 | 145 | 0,394 |
| 4 | 964 | 159 | 0,165 | 741 | 71 | 0,096 | 223 | 88 | 0,395 |
| 5 | 805 | 116 | 0,144 | 670 | 63 | 0,094 | 135 | 53 | 0,393 |
| 6 | 689 | 90 | 0,131 | 607 | 58 | 0,096 | 82 | 32 | 0,390 |
| 7 und später | 599 | 599 | – | 549 | 549 | – | 50 | 50 | – |

Man kann sich diesen Beweis auch durch folgende Überlegung plausibel machen: Die Personen mit hoher Mobilität verlassen zuerst ihre Tätigkeit und lassen die Personen zurück, die eine geringe Mobilität aufweisen. Behandelt man alle Personen gleich (unabhängig von ihrer Mobilitätsrate), dann entsteht der Eindruck eines zeitabhängigen Prozesses mit abnehmender

Karrieremobilität. Bevor man also zeitabhängige Erklärungsmodelle erwägt, sollte man zunächst die Heterogenität des Untersuchungsmaterials kontrollieren. Das kann dadurch geschehen, daß man das Untersuchungsmaterial in homogene Subgruppen zerlegt (z.B. Männer – Frauen) und innerhalb dieser Subgruppen prüft, ob noch eventuelle Zeitabhängigkeiten vorliegen. Sehr viel effizienter ist es jedoch, wenn man die Rate in einem Regressionsmodell von mehreren exogenen Merkmalen (z.B. Geschlecht) abhängig macht. $\mathbf{x}_i$ sei ein Zeilenvektor von Kovariaten, die die Untersuchungseinheit i charakterisieren. In einem allgemeinen Erklärungsmodell sollte daher die Rate zuerst von den Merkmalen der Untersuchungseinheiten abhängen: $r(t|\mathbf{x}_i)$.

### *2.5.2 Zensierung*

Bis jetzt bin ich immer davon ausgegangen, daß der jeweilige Verlaufsprozeß vollständig erhoben wurde, so daß man über Art, Abfolge und Zeitpunkt aller Ereignisse informiert ist. In der Praxis ist der Untersuchungszeitraum jedoch begrenzt, so daß Teile des Veränderungsprozesses unbekannt bleiben (vgl. Abschnitt 1.3). Verlaufsdaten sind in der Regel *rechts*– und manchmal *linkszensiert*. Das Problem linkszensierter Verlaufsdaten existiert dann nicht, wenn der Untersuchungsbeginn $\tau_0$ mit dem Beginn $t_0$ des Prozesses zusammenfällt. Linkszensierte Beobachtungen lassen sich manchmal durch retrospektive Erhebung nacherheben, während rechtszensierte Beobachtungen prinzipiell nicht vervollständigt werden können, da man nicht in die Zukunft blicken kann. Dafür sind die statistischen Probleme bei rechtszensierten Beobachtungen sehr viel einfacher zu lösen als bei Linkszensuren.

Der folgende Abschnitt 2.5.2.1 illustriert die statistischen Probleme, wenn linkszensierte Beobachtungen existieren. Praktikable Alternativen sind nur begrenzt verfügbar, daher wird in der Regel angenommen, daß der Untersuchungsbeginn mit dem Prozeßbeginn zusammenfällt. Abschnitt 2.5.2.2 unterscheidet dann verschiedene Zensierungsmechanismen für den Regelfall rechtszensierter Beobachtungen. Eine wichtige Erkenntnis ist dabei, daß Rechtszensuren wie konkurrierende Risiken betrachtet werden können (vgl. Abschnitt 2.4.1). Schließlich werden in Abschnitt 2.5.2.3 ein paar einfache Schätzer der Rate eines Poisson–Prozesses abgeleitet, die für die empirischen Analysen in den folgenden Kapiteln hilfreich sind.

### 2.5.2.1 Linkszensierte Beobachtungen

Fallen Untersuchungsbeginn $\tau_0$ und Prozeßbeginn $t_0$ nicht zusammen, dann fehlen Informationen über die Anfangsbedingungen des Prozesses zum Zeitpunkt $t_0$ und die Vorgeschichte im Intervall $[t_0,\tau_0)$ vor Eintritt in die Untersuchung. Zu den *Anfangsbedingungen* zählen der Ausgangszustand $z(t_0)$ der Untersuchungseinheiten und die Anfangswerte $\mathbf{x}(t_0)$ der entsprechenden Kovariaten. Zur *Vorgeschichte* $G(t_0,\tau_0)$ rechnet man alle Veränderungen der untersuchten Zustandsvariablen sowie der Kovariaten im Intervall $[t_0,\tau_0)$. Mangelnde Kenntnis sowohl der Anfangsbedingungen als auch der Vorgeschichte kann die Ergebnisse statistischer Analysen erheblich verzerren. Eine formale Ableitung dieser Behauptung würde den Rahmen dieser Arbeit sprengen, daher möchte ich die wesentlichen Überlegungen mit Hilfe eines Beispiels verdeutlichen.

Betrachten wir also einmal folgende Untersuchung über die Mortalität der Bundesbürger: An Hand einer repräsentativen Stichprobe wird festgestellt, wieviel Personen einer Altersklasse innerhalb eines Jahres versterben. Basis dieser Berechnungen ist ein Bevölkerungsquerschnitt der zu einem bestimmten Zeitpunkt $\tau_0$ lebenden Bundesbürger. In der Sprache der Verlaufsanalyse geht es also um das singuläre Ereignis "Tod". Das Geburtsdatum der untersuchten Personen ist ein Beispiel für die Anfangsbedingungen des Prozesses. Zur Vorgeschichte gehört u.a. das Überleben der Bundesbürger bis zum Zeitpunkt der Stichprobenziehung.

Beginnen wir zunächst mit dem zweiten Punkt, dessen Konsequenzen evident sind. Offensichtlich kann die Stichprobe nur die Personen erfassen, die bis $\tau_0$ überlebt haben. Sie vernachlässigt umgekehrt diejenigen, die vorher verstorben sind, und ist somit eine "positive" Auswahl, weil die auf dieser Basis berechnete Mortalität sehr viel geringer ist als sie es wäre, wenn man alle Personen jeder Altersklasse beobachtet hätte. HECKMAN (1979) bezeichnet dies als "sample selection bias", weil die Stichprobe auf der Basis einer Variable (Überleben bis $\tau_0$) ausgewählt wird, die selbst Ergebnis des untersuchten Prozesses ist (daher als endogene Variable bezeichnet). Kann man diese Eigenschaft mit Hilfe anderer Informationen kontrollieren, dann läßt sich die beschriebene Verzerrung korrigieren (vgl. dazu TUMA/HANNAN 1984: 130f.). In der beschriebenen Mortalitätsuntersuchung dürften dafür die Chancen jedoch relativ schlecht stehen, denn um das Überleben bis $\tau_0$ zu prognostizieren, benötigt man Informationen über die Mortalität der einzelnen Altersklassen und diese soll ja erst durch die Untersuchung herausgefunden werden.

Diese Überlegungen gelten im übrigen auch dann, wenn man in der günstigen Situation ist, nicht auf eine solchermaßen selektierte Stichprobe angewiesen zu sein. Angenommen, man möchte die Arbeitslosigkeitsdauer untersuchen und ist dabei nicht auf eine Stichprobe von Arbeitslosen angewiesen, sondern kann auf eine repräsentative Stichprobe von Erwerbspersonen zurückgreifen. Zum Erhebungszeitpunkt wird ein Teil der Untersuchungspersonen bereits arbeitslos sein und die (Fort−)Dauer dieser Arbeitslosigkeiten wird sich ohne Kenntnis ihrer Dauer bis $\tau_0$ und aller vorherigen Arbeitsmarkterfahrungen nur bedingt einschätzen lassen. Man benötigt daher auch bei dieser (nicht selektierten) Stichprobe Informationen darüber, daß die genannten Personen zum Erhebungszeitpunkt arbeitslos sind.

Ein weiteres Problem betrifft die Anfangsbedingungen des Prozesses (HECKMAN 1981). Kehren wir also zu dem Mortalitätsbeispiel zurück: Üblicherweise ist das Geburtsdatum über das Alter der Personen zum Erhebungszeitpunkt bekannt. Wir wollen jedoch einmal annehmen, daß auch diese Information fehlt. Die Berechnung altersspezifischer Mortalitätsziffern ist jetzt natürlich nicht mehr möglich, jedoch läßt sich feststellen, wieviel Personen insgesamt innerhalb des Untersuchungszeitraums versterben. Diese allgemeine Mortalitätsziffer kann jedoch je nach Altersverteilung innerhalb der Stichprobe erheblich verzerrt sein. Ohne Kenntnis der Geburtsdaten, sprich der Anfangsbedingungen, läßt sich diese Verzerrung nicht einschätzen. TUMA und HANNAN (1984: 131f.) zeigen für dieses einfache Beispiel, wie man ohne Kenntnis des Alters jeder einzelnen Untersuchungsperson, aber mit Angaben über die Altersverteilung in der Grundgesamtheit diese unbekannten Anfangsbedingungen kontrollieren kann.

Alle genannten "Reparaturstrategien" können die fehlenden Informationen jedoch nur bedingt ersetzen und sind zudem nicht einfach umzusetzen. Es ist daher immer vorteilhaft, möglichst viele der fehlenden Informationen durch Nacherhebung zu ergänzen. Häufig wird auch einfach der Untersuchungsbeginn mit dem Prozeßbeginn gleichgesetzt, wenn schon keine zeitliche Synchronisation möglich ist. In anderen Fällen beschränkt man die Auswertung auf die Wartezeiten, die innerhalb des Untersuchungszeitraums neu beginnen (z.B. eine Analyse der Neuzugänge in Arbeitslosigkeit statt des Bestandes zum Zeitpunkt der Stichprobenziehung). Für die beiden folgenden Abschnitte möchte ich davon ausgehen, daß der Prozeß von Anfang an beobachtet wurde und Linkszensuren nicht auftreten.

### 2.5.2.2 Rechtszensierte Beobachtungen

Angenommen, man beobachtet einen Verlaufsprozeß von Prozeßbeginn $\tau_0 = t_0$ bis zum Zeitpunkt $\tau_e$ – insgesamt also $\tau = \tau_e - \tau_0$ Zeiteinheiten: Wenn wir uns der Einfachheit halber zunächst auf singuläre Ereignisse beschränken, dann hat innerhalb dieses Untersuchungszeitraums $[t_0, \tau_e]$ ein Teil der Untersuchungseinheiten ein Ereignis zum Zeitpunkt $t_i$ ($t_i < \tau$) und ein anderer Teil wird am Untersuchungsende zensiert mit Überlebenszeit $t_i = \tau$. Wie in Abschnitt 1.3 beschrieben, handelt es sich hierbei um *Typ I zensierte Beobachtungen* mit einer für alle Untersuchungseinheiten gleichen *fixen Untersuchungsdauer.*

Die Ausgangsdaten sind also im Gegensatz zu unseren bisherigen Überlegungen zweigeteilt: Ein Teil der Stichprobe überlebt ohne Ereignis bis $\tau$. Die Wahrscheinlichkeit dafür ist $S(\tau)$. Für diese Teilgruppe ist die Wartezeit keine Zufallsvariable, sondern eine feste Größe entsprechend der Untersuchungsdauer. Ein anderer Teil hat ein Ereignis. Die Wahrscheinlichkeit dafür ist

$$P(t_i | t_i < \tau) = \frac{f(t_i)dt}{1 - S(\tau)} \tag{2.48}$$

Wenn man die Unterscheidung zwischen beiden Teilgruppen in einer entsprechenden (Status–)Variablen $\delta_i$ festhält ($\delta_i = 1$ Ereignis, $\delta_i = 0$ Zensierung), dann sind nunmehr zwei Zufallsvariablen zu betrachten: $\delta$ (Ereignis oder Zensierung) sowie T (Wartezeit, aber nur für Ereignisse). Die Wahrscheinlichkeit beider Zufallsvariablen ist entweder $P(t_i = \tau, \delta_i = 0) = S(\tau)$ oder $P(t_i, \delta_i = 1) = P(t_i | t_i < \tau) P(t_i < \tau) = f(t_i)dt$. Ihre gemeinsame Verteilung kann mit Hilfe der Variablen $\delta_i$ und unter Verwendung von (2.14) elegant wie folgt zusammengefaßt werden:

$$f(t_i, \delta_i) = f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} = r(t_i)^{\delta_i} S(t_i) \tag{2.49}$$

Da alle Untersuchungseinheiten zum gleichen Zeitpunkt zensiert werden, kann man davon ausgehen, daß Zensierung und die eigentlich interessierenden Ereignisse voneinander unabhängig sind.

Ein allgemeinerer Zensierungsmechanismus ergibt sich, wenn die Beobachtungsdauer $\tau_i$ selbst eine Zufallsvariable ist, weil beispielsweise die Untersuchungseinheiten zu unterschiedlichen Zeitpunkten in die Untersuchung eintreten (vgl. Abb. 1.4c) oder aus anderen Gründen unterschied-

lich lange beobachtet werden können. In diesem Fall spricht man von *Typ I zensierten Beobachtungen* mit einer *zufälligen Untersuchungsdauer*. Die gemeinsame Verteilung der beiden Zufallsvariablen T und $\tau$ kann in Analogie zum Modell konkurrierender Risiken entwickelt werden, wobei Zensierung als konkurrierendes Risiko neben dem eigentlich interessierenden Risiko betrachtet wird. Unterstellen wir also zwei voneinander unabhängige Zeitdauern, die Wartezeit $T_i$ bis zum Eintritt des Ereignisses und die Zeitdauer $\tau_i$ bis zum Zeitpunkt der Zensierung. Die beobachtete Wartezeit $t_i$ entspricht dann der kleineren $t=\min(T_i,\tau_i)$ der beiden Zeitdauern. Wenn alle Untersuchungseinheiten im gleichen Ausgangszustand j beginnen, dann lassen sich also zwei mögliche Wechsel unterscheiden: einmal in den Zustand k=d (Ereignis) und das andere Mal nach k=c (Zensierung). $r_{jd}(t)$ und $r_{jc}(t)$ seien die entsprechenden Übergangsraten und unter Verwendung von (2.41) gilt für Ereignisse

$$(2.50)\qquad f^*_{jd}(t) = r_{jd}(t)\exp\left[-\int_0^t (r_{jd}(u)+r_{jc}(u))du\right]$$
$$= r_{jd}(t)\,S_{jd}(t)\,S_{jc}(t)$$

bzw. für Zensierungen

$$(2.51)\qquad f^*_{jc}(t) = r_{jc}(t)\exp\left[-\int_0^t (r_{jd}(u)+r_{jc}(u))du\right]$$
$$= r_{jc}(t)\,S_{jc}(t)\,S_{jd}(t)$$

Dies kann wiederum unter Verwendung von $\delta_i$ kombiniert werden zu

$$(2.52)\qquad f^*_{jk}(t) = [r_{jd}(t)\,S_{jd}(t)\,S_{jc}(t)]^{\delta_i}\,[r_{jc}(t)\,S_{jc}(t)\,S_{jd}(t)]^{1-\delta_i}$$
$$= [r_{jd}(t)^{\delta_i}\,S_{jd}(t)]\,[r_{jc}(t)^{1-\delta_i}\,S_{jc}(t)]$$

Der erste Term beschreibt das Auftreten von Ereignissen, der zweite das Auftreten von Zensierungen. Wenn die obige Unabhängigkeitsannahme richtig ist, dann kann man zeigen, daß die Rate $r_{jd}(t)$, mit der die eigentlich interessierenden Ereignisse eintreten, unabhängig von der Zensierungsrate $r_{jc}(t)$ geschätzt werden kann, d.h. der zweite Term kann bei der Schätzung vernachlässigt werden. Zu prüfen ist allerdings im Einzelfall, ob tatsächlich ein unabhängiger Zensierungsmechanismus vorliegt. Man beachte des wei-

teren, daß (abgesehen von den Indizes) der erste Term und (2.49) identisch sind. Wie LAWLESS (1982: 38ff.) zeigt, kann der Ausdruck $r_{jd}(t)^{\delta_i}S_{jd}(t)$ für rechtszensierte Prozesse mit Zensierungsmechanismen unterschiedlicher Art verwendet werden.

### 2.5.2.3 Einfache Schätzer für rechtszensierte Poisson – Prozesse

Gleichung (2.17c) zeigt die durchschnittliche Zustandsdauer E(t) eines zeitkonstanten Prozesses. Wäre diese durchschnittliche Dauer bekannt, könnte man umgekehrt aus 1/E(t) die Rate $\lambda$ berechnen. Dabei wird jedoch implizit davon ausgegangen, daß der gesamte Prozeß von Anfang $t=0$ bis Ende $t=\infty$ beobachtet wird (daher das Integral von 0 bis $\infty$). Häufig wird ein Prozeß jedoch nur $\tau$ Zeiteinheiten erhoben und ein Teil der Beobachtungen ist rechtszensiert. In diesem Abschnitt möchte ich mich mit der Frage beschäftigen, ob es auch für diesen realistischen Anwendungsfall einen einfachen Schätzer der Rate gibt. Ich übernehme dazu die Annahmen des vorherigen Abschnitts (Typ I zensierte Beobachtungen bei fixer Untersuchungsdauer $\tau$) und beschränke mich auf zeitkonstante Prozesse mit singulären Ereignissen.

N sei die Gesamtzahl aller Untersuchungseinheiten, von denen $D_1$ innerhalb des Zeitraums $[0,\tau]$ ein Ereignis aufweisen, während bei C kein Ereignis beobachtet werden kann. $U_1$ sei die Summe der Zeitdauern bis zum ersten Ereignis der $D_1$ Untersuchungseinheiten mit mindestens einem Ereignis. Die Dauer jeder zensierten Beobachtung entspricht der Untersuchungsdauer $\tau$. Insgesamt seien es $C\tau=V$ Zeiteinheiten. Mit diesen Ausgangsdaten lassen sich drei verschiedene Durchschnittswerte berechnen:

1. die durchschnittliche Dauer bis zu einem Ereignis bezogen auf alle Untersuchungseinheiten

   (2.53) $$\bar{t}_1 = U_1 / N$$

2. die durchschnittliche Dauer bis zu einem Ereignis bezogen auf alle Untersuchungseinheiten mit mindestens einem Ereignis

   (2.54) $$\bar{t}_2 = U_1 / D_1$$

3. die durchschnittliche Dauer aller unzensierten und zensierten Wartezeiten bezogen auf alle Untersuchungseinheiten

$$\bar{t}_3 = U_1 / N + C\tau / N = (U_1 + V) / N \tag{2.55}$$

(2.55) legt im übrigen folgende Verallgemeinerung auf wiederholbare singuläre Ereignisse nahe: Wenn man sich überlegt, daß eine Aneinanderreihung aller $l=1,2,\ldots$ Zustandsdauern $t_l$ einer Untersuchungseinheit genau die Untersuchungsdauer $\tau$ ergibt, dann ist $N\tau$ die Summe aller Wartezeiten (inkl. Wiederholungen) über alle Untersuchungseinheiten. Bezogen auf alle (erste und folgende) Ereignisse D ergibt sich schließlich ein vierter Durchschnittswert

$$\bar{t}_4 = \frac{N\tau}{D} = \frac{U + V}{D} \tag{2.56}$$

U ist dabei die Summe aller unzensierten Wartezeiten und V die Summe der zensierten Wartezeiten. Zensierte Wartezeiten entsprechen $\tau$ für Untersuchungseinheiten ohne Ereignis bzw. der Dauer der letzten Wartezeit für Einheiten mit mindestens einem Ereignis.

Im nächsten Schritt wäre nun zu fragen, ob einer der vier Werte eine einfache Funktion der zeitkonstanten Rate $\lambda$ ist. Dies würde es erlauben, mit $\bar{t}_1$, $\bar{t}_2$, $\bar{t}_3$ oder $\bar{t}_4$ die Rate $\lambda$ zu schätzen. Alle vier Funktionen lassen sich mit den bisher erworbenen Kenntnissen über Erwartungswerte ableiten, wenn man die begrenzte Untersuchungsdauer berücksichtigt. Geht es also um den Erwartungswert aller Wartezeiten, die kleiner als die Untersuchungsdauer $\tau$ sind (vgl. 2.53), dann ist in (2.17c) von 0 bis $\tau$ zu integrieren:

$$E(t|t<\tau) = \int_0^{\tau} t\, f(t)dt = \lambda^{-1} [1 - (\lambda\tau+1) \exp(-\lambda\tau)] \tag{2.57}$$

Er ist natürlich kleiner als der Erwartungswert (2.17c), da nur ein Teil der Stichprobe bis zum Zeitpunkt $\tau$ ein Ereignis hat. Soll dieser Erwartungswert nicht für alle N Untersuchungseinheiten, sondern nur für die mit Ereignis berechnet werden (vgl. 2.54), ist entsprechend durch die Wahrscheinlichkeit, ein Ereignis innerhalb $[0,\tau]$ zu haben, zu teilen:

$$(2.58) \qquad E(t|\delta_i=1, t<\tau) = \frac{\int_0^\tau t\, f(t)\, dt}{\int_0^\tau f(t)\, dt}$$

$$= \lambda^{-1} \frac{1 - (\lambda\tau+1)\exp(-\lambda\tau)}{1 - \exp(-\lambda\tau)}$$

Beide Erwartungswerte (2.57) und (2.58) haben jedoch den Nachteil, eine äußerst komplizierte Funktion des Parameters $\lambda$ und der Untersuchungsdauer $\tau$ zu sein (man versuche einmal (2.57) oder (2.58) nach $\lambda$ aufzulösen). Daher kann man sie kaum verwenden, um umgekehrt aus den (empirischen) Durchschnittswerten $\bar{t}_1$ oder $\bar{t}_2$ das Ereignisrisiko zu berechnen.

Betrachten wir jedoch den Durchschnitt $\bar{t}_3$ aller ersten Tätigkeitsdauern mit und ohne Wechsel (vgl. 2.55). Der Erwartungswert der Summe der zensierten und unzensierten Wartezeiten entspricht nämlich folgendem Ausdruck:

$$(2.59) \qquad E(u+v|\tau) = \int_0^\tau t\, f(t)\, dt + \int_\tau^\infty \tau\, f(t)\, dt$$

$$= \lambda^{-1}[1 - (\lambda\tau+1)\exp(-\lambda\tau)] + \tau\exp(-\lambda\tau)$$

$$= \lambda^{-1}[1 - \exp(-\lambda\tau)]$$

$$= \lambda^{-1}[1 - S(\tau)]$$

$1-S(\tau)$ entspricht dem Anteil derjenigen, die im Untersuchungszeitraum mindestens ein Ereignis aufweisen: $1-S(\tau)=D_1/N$. Durch Gleichsetzen von (2.55) und (2.59) und durch Auflösen nach $\lambda$ erhält man einen einfachen Schätzer für das Ereignisrisiko:

$$(2.60) \qquad \hat{\lambda} = \frac{1 - S(\tau)}{t_3} = \frac{\frac{D_1}{N}}{\frac{U_1 + V_1}{N}} = \frac{D_1}{U_1 + V_1}$$

Er entspricht dem Verhältnis aller ersten Ereignisse bezogen auf alle (zensierten und unzensierten) ersten Zustandsdauern. Das ist genau der Ausdruck, den man sich gemeinhin unter einer Rate vorstellt, nämlich die Anzahl der Ereignisse bezogen auf den Zeitraum, in dem sie eingetreten

sind oder hätten eintreten können. Die gleichzeitige Verwendung von zensierten und unzensierten Beobachtungen macht deutlich, daß man bei der Analyse von Verlaufsdaten unvollständig erhobene Beobachtungen nicht vernachlässigen sollte.

Jetzt erkennt man auch, daß der Kehrwert des vierten Durchschnittswerts $\bar{t}_4$ nur eine Verallgemeinerung des Schätzers (2.60) auf alle im Untersuchungszeitraum auftretenden Ereignisse ist. Er kann auch aus unseren Überlegungen zu Modellen mit wiederholbaren Ereignissen abgeleitet werden. Nach (2.45) entspricht der Parameter $\lambda$ dem Verhältnis aller Ereignisse D bezogen auf das $N-$fache der Untersuchungsdauer.

# 3. *Anwendungsbeispiele und EDV – technische Umsetzung von Verlaufsanalysen*

In den folgenden Kapiteln dieser Arbeit möchte ich mich mit der Frage beschäftigen, wie man einen konkreten Verlaufsdatensatz auswertet, um entweder Hypothesen über den zugrundeliegenden Veränderungsprozeß zu generieren oder um schon vorhandene Hypothesen an Hand empirischer Daten zu überprüfen. Die praktische Relevanz der Verlaufsdatenanalyse, ihre Möglichkeiten aber auch ihre Grenzen lassen sich am besten an Hand von Beispielen erläutern. Exemplarisch für viele andere Fragestellungen will ich bei meinen folgenden Ausführungen zwei Datensätze verwenden, die in einem Fall die Bewährungszeit einer Gruppe von Strafentlassenen und im anderen Fall berufliche Mobilitätsprozesse erfassen. In den Abschnitten 3.1 und 3.2 möchte ich kurz die Herkunft der Daten sowie die wesentlichen Hypothesen der Untersuchung darstellen, um schließlich im letzten Abschnitt 3.3 ausführlich auf das Datenmanagement von Verlaufsdaten einzugehen. Dabei geht es um die Frage, wie Verlaufsdaten sinnvollerweise aufbereitet werden sollen, um mit gängiger Software ausgewertet zu werden. Eine Übersicht über vorhandene Programmpakete und die dort verfügbaren statistischen Prozeduren zur Verlaufsdatenanalyse rundet die Darstellung ab.

## *3.1 Strafentlassenen – Daten*

Die Relevanz zeitbezogener Fragestellungen zeigt sich bei vielen sozialwissenschaftlichen Anwendungsproblemen. Dies gilt ganz besonders für die Evaluierung sozialer und politischer Programme (vgl. die Diskussion in der Einleitung, insbesondere Abschnitt 1.1.2). Bewährungshilfe im Strafvollzug läßt sich als ein solches Programm verstehen, das die Resozialisation von Strafentlassenen zum Ziel hat. An Hand der Daten, die ein Bewährungshelfer über die von ihm betreuten Personen gesammelt hat, habe ich mich mit der Frage beschäftigt, von welchen Faktoren es abhängt, ob ein Strafentlassener innerhalb der Bewährungszeit wieder straffällig wird, und ob dieses

Rückfallrisiko im Zeitablauf variiert (ANDRESS 1984a).[1] Tabelle 3.1 zeigt die wesentlichen Variablen dieses Datensatzes (fortan als KFN – Daten abgekürzt).

*Tabelle 3.1: Daten über 64 Strafentlassene*

| Variable | Kodierung | Min | Max | Mittel-wert | Standard-abweichung |
|---|---|---|---|---|---|
| Person | Nummer[a)] | 1 | 65 | 33,48 | 18,65 |
| Haftgrund | 1 = Diebstahl, 0 = Rest | 0 | 1 | 0,48 | 0,50 |
| Bewährungszeit | Monate | 2 | 37 | 23,98 | 10,51 |
| Rückfällig? | 1 = ja, 0 = nein | 0 | 1 | 0,48 | 0,50 |
| Datum der Straftat | Monat | 0 | 21 | 4,31 | 6,00 |
| Termine | 1 = ja, 0 = nein | 0 | 1 | 0,23 | 0,43 |
| Schulden | 1 = ja, 0 = nein | 0 | 1 | 0,59 | 0,50 |
| Arbeit | 1 = ja, 0 = nein | 0 | 1 | 0,19 | 0,39 |
| Soziale Probleme | 1 = ja, 0 = nein | 0 | 1 | 0,28 | 0,45 |
| Wohnung | 1 = ja, 0 = nein | 0 | 1 | 0,09 | 0,29 |
| Wartezeit | Monate | 2 | 37 | 17,03 | 11,58 |

a) N = 64, eine Person bleibt wegen fehlender Werte unberücksichtigt

Von den insgesamt 64 Personen sind fast die Hälfte (N = 31) innerhalb der Bewährungszeit rückfällig. Die Erfolgschancen der Bewährungshilfe dürften dabei sehr stark von den sonstigen sozialen Belastungen der Person abhängen. Der betreuende Sozialarbeiter hat in diesem Zusamenhang festgehalten, ob sein Mandant

a) vereinbarte Termine nicht einhielt,
b) keinen Eigenbeitrag zur Schuldenregulierung lieferte,
c) keiner Beschäftigung nachging,
d) Probleme in der Familie oder im sozialen Umfeld hatte oder
e) keine eigene Wohnung besaß.

Anzunehmen ist, daß das Rückfallrisiko dann sehr viel höher ist, wenn in einem oder mehreren dieser Bereiche zusätzliche Schwierigkeiten für den Betroffenen auftreten. Darüber hinaus ist zu erwarten, daß sich direkt nach Haftentlassung viele Umstellungsprobleme kumulieren und daher Anpassungsschwierigkeiten und Gefährdungen zu Beginn am größten sind. Wie

1) Die Daten wurden freundlicherweise von Herrn B. Hesener (Kriminologisches Forschungsinstitut Niedersachsen e.V., Hannover) zur Verfügung gestellt.

zuvor angedeutet, ist also davon auszugehen, daß das Rückfallrisiko im Zeitablauf variiert.

Das interessierende Ereignis der folgenden Verlaufsanalyse ist die eventuelle erneute Straftat der untersuchten Personen – also ein singuläres Ereignis.[1] Jede Person wird nur eine begrenzte Zeit (für die Dauer der Bewährung) beobachtet. Falls innerhalb dieses Zeitraums keine Straftat begangen wird, heißt das nicht, daß die Person nie mehr rückfällig wird – es liegen uns nur über diese eventuellen "späteren" Straftaten keine Informationen vor. Die *Wartezeit* bis zum interessierenden *Ereignis* besteht also aus dem Zeitpunkt der Strafttat, falls die Person innerhalb der Bewährungszeit rückfällig wird. Ansonsten wird die Wartezeit nach Ablauf der Bewährungsfrist *zensiert*.

Eine methodische Schwierigkeit ergibt sich eventuell dadurch, daß Zensierungen (Bewährung) und Ereignisse (Straftat) nicht unabhängig voneinander sind, wenn Personen mit höherem Rückfallrisiko eine längere Bewährungszeit haben. Man könnte etwa vermuten, daß typische Rückfalltäter höhere Haftstrafen und damit auch längere Bewährungszeiten auferlegt bekommen. Diese Frage läßt sich jedoch mit den vorliegenden Daten nur sehr schwer beantworten, zumal der statistischen Überprüfung der Unabhängigkeitsannahme enge Grenzen gesetzt sind. Ich gehe daher vereinfachend von der Gültigkeit dieser Annahme aus.

## *3.2 Mikrozensus – Zusatzerhebung 1971*

Die Daten stammen aus der Mikrozensus – Zusatzerhebung aus dem Jahr 1971 über berufliche und soziale Umschichtung der Bevölkerung (fortan als MZ71 – Daten abgekürzt).[2] Aus der Gesamtstichprobe von ca. 450.000 Personen wurden die männlichen Befragten ausgewählt, die in den Jahren 1961 bis 1963 in das Erwerbsleben eingetreten sind – eine sogenannte *Berufsanfängerkohorte* (zu den Einzelheiten der Stichprobenziehung und Datenaufbereitung vgl. ANDRESS 1984b).

---

1) Wiederholbare Ereignisse (mehrfache Straftaten innerhalb der Bewährungszeit) sollen in dieser Analyse außer Acht gelassen werden.

2) TEGTMEIER (1976) informiert über die Einzelheiten der Zusatzerhebung. Die eigentliche Auswertung der Daten erfolgte im Rahmen des VASMA – Projektes (Universität Mannheim). Prof. Dr. W. Müller sei dafür an dieser Stelle gedankt.

Das Design der Zusatzerhebung sah eine retrospektive Erhebung der Lebensläufe für den Zeitraum 1960–1971 vor, wobei sowohl erwerbsbezogene als auch nicht erwerbsbezogene Tätigkeiten (Ausbildung, Arbeitslosigkeit, Rentner/Pensionäre, Hausarbeit, Sonstiges) erfaßt werden sollten. Die Befragten wurden zunächst aufgefordert, ihre Tätigkeit im Jahre 1960 anzugeben, und mußten dann alle folgenden Tätigkeitsänderungen (jahresweise) auflisten. Schließlich wurden sie gefragt, welche Tätigkeit sie zum Zeitpunkt der Erhebung (April 1971) ausübten. Bei jeder Angabe wurde das entsprechende Jahr sowie Tätigkeit und Wirtschaftszweig der Tätigkeit vermerkt. Das entsprechende Erhebungsschema unterschied zwischen 15 verschiedenen Wirtschaftszweigen und 43 verschiedenen Tätigkeiten, wobei letztere nach sozialrechtlicher Stellung, Qualifikation, Anforderungsgrad, Dispositionsbefugnis und Betriebsgröße (bei Selbständigen) differenziert wurden (ANDRESS 1984b: 294). "In Kombination mit Angaben zum Wirtschaftszweig konnte auf diese Weise die berufliche Situation des Befragten, seiner Eltern oder seines Ehegatten ausreichend charakterisiert werden" (TEGTMEYER 1976: 9).

Aus diesen Informationen wurden für die Analyse der ersten zehn Berufsjahre Verlaufsdaten generiert, wie sie beispielhaft in Tabelle 1.1 bereits vorgestellt wurden. Veränderungen des (beruflichen) Lebenslaufs werden durch das Erhebungsschema der Zusatzerhebung aber nur soweit erfaßt, wie sie sich in den o.g. Kategorien niederschlagen. *Tätigkeiten* sind nach dieser Definition ein konkretes Kategorienpaar der 43 verschiedenen Tätigkeiten bzw. 15 verschiedenen Wirtschaftszweige. Tätigkeitswechsel sind die Veränderungen des beruflichen Lebenslaufs, die zu einer Änderung dieses Kategorienpaares führen.[1] Ausdrücklich sei darauf hingewiesen, daß die solchermaßen definierten Tätigkeiten und Tätigkeitswechsel nicht alle tatsächlich stattfindenden Tätigkeitsänderungen erfassen – z.B. dann nicht, wenn ein Kfz–Mechaniker aus der Reparaturabteilung eines Betriebes in das Ersatzteillager eines anderen Betriebes wechselt, weil er in beiden Fällen als Facharbeiter im verarbeitenden Gewerbe erfaßt wird.

Insgesamt umfaßt die ausgewählte Kohorte 10.666 Männer mit vollständigen Angaben über ihren beruflichen Lebenslauf, die innerhalb der ersten

1) Ausbildung und Wehrdienst wurden dabei nicht als eigenständige Tätigkeiten, sondern als Unterbrechung der Berufslaufbahn gewertet. Die Anzahl unterschiedener Tätigkeiten reduziert sich damit auf 43 – 13 = 30.

zehn Berufsjahre 18.094 verschiedene Tätigkeiten ausgeübt haben.[1] 5.288 Männer zeigen überhaupt keine Veränderungen innerhalb des Untersuchungszeitraumes, 5.378 haben zum Teil mehrmals ihre Tätigkeit gewechselt. Alle Tätigkeiten ohne Wechsel, d.h. die 5.288 Tätigkeiten der "immobilen" Männer und die 5.447 Tätigkeiten, die die "mobilen" Männer *zuletzt* ausgeübt haben, sind *zensierte Beobachtungen*. Tätigkeiten, die mit einem Wechsel enden, sind dagegen *Beobachtungen mit Ereignis*. Mit Hilfe eines Statusindex von HANDL (1977) wurden diese Wechsel nach sozialen Auf- und Abstiegen bzw. Wechseln ohne wesentliche Veränderungen des sozialen Status unterschieden (horizontale Mobilität bei Differenzen $\leq \pm 3$ Indexpunkte). Im Gegensatz zu den Strafentlassenen-Daten handelt es sich hier also um einen Prozeß, in dem *multiple Ereignisse* möglich sind, die sich zudem *wiederholen* können. Die folgende Tabelle 3.2 zeigt die Verteilung aller ersten Tätigkeiten nach Art des Wechsels.

*Tabelle 3.2: Dauer erster Tätigkeiten nach Art des Wechsels*

| Jahre | Abstieg | Aufstieg | horizontale Mobilität | ohne Wechsel | Insgesamt N | Insgesamt % |
|---|---|---|---|---|---|---|
| < 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1-2 | 129 | 223 | 76 | 0 | 428 | 4,0 |
| 2-3 | 126 | 306 | 92 | 0 | 524 | 4,9 |
| 3-4 | 170 | 354 | 90 | 0 | 614 | 5,8 |
| 4-5 | 114 | 364 | 70 | 0 | 548 | 5,1 |
| 5-6 | 119 | 368 | 61 | 0 | 548 | 5,1 |
| 6-7 | 99 | 397 | 49 | 0 | 545 | 5,1 |
| 7-8 | 125 | 409 | 73 | 0 | 607 | 5,7 |
| 8-9 | 143 | 398 | 130 | 0 | 671 | 6,3 |
| 9-10 | 115 | 251 | 118 | 1638 | 2122 | 19,9 |
| 10-11 | 65 | 172 | 172 | 1600 | 2009 | 18,8 |
| > 11 | 0 | 0 | 0 | 2050 | 2050 | 19,2 |
| Insges. | 1205 | 3242 | 931 | 5288 | 10666 | 100,0 |

Wie zuvor die Untersuchung über Strafentlassene ist auch dieser Datensatz nicht ohne methodische Probleme. Auf Grund der jahresweisen Erfassung von Tätigkeitsänderungen ist er ein typisches Beispiel einer *ungenauen*

1) Auf Grund fehlender Werte bei den Kovariaten können sich diese Fallzahlen weiter verringern.

*Erhebung von Verlaufsdaten.*[1] Im Ergebnis unterscheidet er sich nicht wesentlich von zeitdiskret erhobenen Verlaufsdaten. Man kann auf diese Problematik unterschiedlich reagieren (vgl. Abschnitt 1.3):

1. Man ignoriert die Ungenauigkeiten und behandelt die Zeitangaben wie exakt gemessene Wartezeiten.
2. Man berücksichtigt bei der Schätzung zeitkontinuierlicher Modelle, daß die Daten quasi gruppiert erhoben wurden.
3. Man verwendet gleich ein zeitdiskretes Modell.

Für die Zwecke dieser Einführung möchte ich zunächst von der ersten Strategie Gebrauch machen. Sie läßt sich mit Ergebnissen von Monte-Carlo-Studien rechtfertigen, die zeigen, daß sich die Ungenauigkeit der Messungen nicht gravierend auf die Schätzung der Modellparameter auswirkt (CARROL et al. 1978, TUMA et al. 1979: 850, kritisch jedoch ARMINGER 1984c, GALLER 1986). Soweit es im Rahmen dieser Einführung möglich ist, soll diese Ad-hoc-Strategie jedoch durch angemessenere Modelle und Schätzverfahren kontrolliert werden. Es ist anzunehmen, daß die substantiellen Unterschiede geringfügig sind.

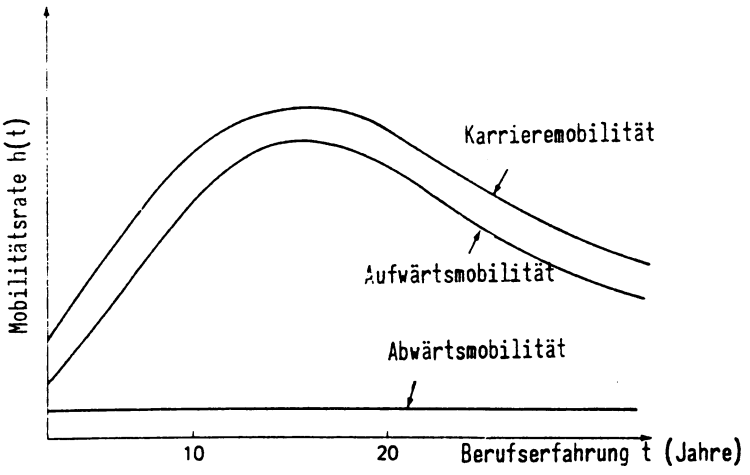
Das wesentliche Ziel meiner Analyse dieses Datensatzes war die Evaluation unterschiedlicher Muster intragenerationeller Mobilität (ANDRESS 1983). Meine Annahme war, daß Auf- und Abstiege auf unterschiedliche Art und Weise durch Merkmale der Person, der jeweiligen Tätigkeit sowie der ökonomischen Rahmenbedingungen beeinflußt werden. Ich bin auch davon ausgegangen, daß Auf- und Abstiege im Verlauf eines Berufslebens unterschiedlich verteilt sind. Ähnliche Untersuchungen wurden von TUMA und SÖRENSEN mit amerikanischen Daten durchgeführt (TUMA 1976, 1982a; SÖRENSEN/TUMA 1978).

Die Grundidee meines Ansatzes ist die Annahme, daß Abstiege mehr als Aufstiege von Faktoren abhängen, die außerhalb des Einflusses einer Person liegen. Konkret gesprochen gehe ich davon aus, daß Abstiege viel mehr von ökonomischen Randbedingungen (Verfügbarkeit von Stellen, schrumpfende Industrien, Entlassungen, Wirtschaftszyklus) abhängen als von persönlichen Merkmalen. Während einer Periode wirtschaftlichen Wachstums wie in dem hier untersuchten Zeitraum von 1960 bis 1971 treten solche externen Einflüsse eher zufällig im Zeitablauf auf. Von daher

1) Die jahresweise Erfassung erklärt im übrigen, daß in Tabelle 3.2 keine Tätigkeit im Intervall [0,1) endet. Jede Tätigkeit ist mindestens ein Jahr lang.

ist anzunehmen, daß bei einer Kohorte von Berufsanfängern, die in dieser Zeit in den Arbeitsmarkt eintritt, das Abstiegsrisiko konstant und in den ersten Berufsjahren sehr niedrig ist.

*Abbildung 3.1: Karrieremobilität und Berufserfahrung*




Im Gegensatz dazu ist es wahrscheinlich, daß die Karrieremobilität im allgemeinen und die Aufwärtsmobilität im besonderen in den ersten Berufsjahren zunehmen, um nach dem Erreichen einer stabilen Berufsposition allmählich auf einen geringeren Wert abzusinken. Diese Hypothesen sind in Abbildung 3.1 graphisch dargestellt: Während Mobilitäts- und Aufstiegsrate in den ersten 10 bis 15 Berufsjahren kontinuierlich zunehmen, später aber wieder absinken, ist das Risiko eines Abstiegs im Zeitablauf konstant. Da der Untersuchungszeitraum der MZ-Zusatzerhebung nur die ersten 10 Berufsjahre der ausgewählten Berufsanfängerkohorte abdeckt, wird die Umkehrung der Mobilitätsrate leider nicht mehr erfaßt. Generell erwartet man jedoch für diesen begrenzten Zeitraum einen positiven Effekt der Berufserfahrung (gemessen in Jahren seit Berufseintritt) auf die Mobilitäts- und Aufstiegsrate, während ein solcher Einfluß bei der Abstiegsrate praktisch zu vernachlässigen ist.

Ökonomische Randbedingungen machen sich jedoch nicht nur durch eine unterschiedliche zeitliche Verteilung verschiedener Arten von Tätigkeitswechseln bemerkbar. Sie sind auch direkt meßbar und können mit

anderen, z.B. persönlichen Merkmalen verglichen werden. Tabelle 3.3 zeigt dazu die wesentlichen Variablen des Datensatzes und deren vermutete Einflüsse. Ich will sie an dieser Stelle nur kurz kommentieren. Eine ausführliche theoretische Ableitung sowie eine Diskussion ihrer Operationalisierung findet sich in ANDRESS (1984b).

Ich gehe davon aus, daß der Status des Vaters einen positiven Effekt auf die Aufwärtsmobilität hat, da Forschungsarbeiten zur Intergenerationenmobilität in der Regel eine positive Korrelation zwischen sozialer Herkunft und Statuszuweisung ergeben. Ausbildung ist bekanntlich eine Investition in Humankapital, die vor Dequalifikation schützt: Je besser also die Ausbildung einer Person ist, um so größer ist die Chance eines Aufstiegs und um so geringer das Risiko eines Abstiegs. Ich nehme auch an, daß alle Personen bei Berufseintritt unterbewertet werden, so daß Aufstiege bei den ersten Tätigkeitswechseln wahrscheinlicher sind und in dem Maße abnehmen, in dem Personen in statusmäßig höheren Positionen beschäftigt werden. Wachsende Industriebranchen (gemessen an der Beschäftigtenzahl) bieten sehr viel mehr Beschäftigungschancen, so daß ein Wechsel in diese Branchen eine ausgezeichnete Möglichkeit darstellt, einen besseren Job zu erhalten: Aufstiege sind also wahrscheinlicher als Abstiege und vice versa. Umgekehrt wäre zu überprüfen, ob eine Person, die einen wachsenden Wirtschaftszweig verläßt (die also praktisch ihre Chance nicht wahrnimmt), sehr viel eher einen Abstieg erfährt. In den sechziger Jahren war in der BRD die Nachfrage nach Arbeit sehr viel höher als das Arbeitsangebot, ausgenommen im Rezessionsjahr 1967. Die Aufstiegschancen waren also allgemein sehr gut und soziale Abstiege sollten daher sehr selten auftreten. Lediglich bei Tätigkeitswechseln, die im Rezessionsjahr 1967 stattfanden, kann man vermuten, daß das Abstiegsrisiko höher war.

Außerberufliche Ereignisse, die sonst noch im Leben einer Person stattfinden, sind von ganz besonderem Interesse für die Analyse intragenerationeller Mobilität, da man relativ wenig Erkenntnisse darüber hat, wie familiäre, gesundheitliche und sonstige Einflüsse den beruflichen Lebenslauf verändern. Ich untersuche in diesem Zusammenhang den Einfluß des Wehrdienstes und die Teilnahme an Fort– und Weiterbildungsmaßnahmen. Fort– und Weiterbildung sind eng verknüpft mit beruflichem Aufstieg, so daß ihr Einfluß auf die Karrieremobilität eigentlich noch sehr viel größer sein sollte als der der Ausbildung, auf jeden Fall aber in die gleiche Richtung verlaufen sollte. Ich messe Fort– und Weiterbildung durch eine Dummy–Variable, die angibt, ob eine Tätigkeit zwecks Teilnahme an einer Fort– oder Weiterbildungsveranstaltung unterbrochen wurde. In

*Tabelle 3.3: Variablen der MZ – Zusatzerhebung – Kodierung und Hypothesen*

| Variablengruppe | Variable (Variablenname) | Kodierung | Erwarteter Effekt auf Aufwärts-mobilität | Abwärts-mobilität | Mittelwert N = 13075 (N = 7587) | Standard-abweichung |
|---|---|---|---|---|---|---|
| zeitkonstante persönliche Merkmale | Status des Vaters (vater) | sozioökonomischer Status nach HANDL (1977) | erhöht | ? | 117,4 (115,1) | 90,4 (90.0) |
| | Ausbildung des Befragten (ausb) | Dauer der Schul- und Berufsausbildung in Jahren nach MÜLLER (1977) | erhöht | verringert | 11,3 (11,3) | 2,3 (2,3) |
| Tätigkeitsmerkmale | Status (status) | sozioökonomischer Status (HANDL 1977) der Tätigkeit | verringert | ? | 116,3 (102,7) | 86,7 (78,7) |
| | Abfolge der Tätigkeiten (taetnr) | 1 = erste, 2 = zweite, 3 = alle folgenden Tätigkeiten | verringert | ? | 1,5 (1,0) | 0,7 (0,0) |
| wirtschaftliche Rahmenbedingungen | Beschäftigtenwachstum Ausgangsbranche (ausgbesh) | Beschäftigtenwachstum 1961 – 70 (ANDRESS 1984b) des Wirtschaftszweiges, in dem die Tätigkeit ausgeübt wird | ? | erhöht | 115,2 (113,4) | 38,7 (39,0) |
| | Beschäftigtenwachstum Zielbranche (zielbesh) | Beschäftigtenwachstum 1961 – 70 (ANDRESS 1984b) Wirtschaftszweig nach Wechsel | erhöht | ? | 117,7 (117,0) | 38,3 (38,6) |
| | Tätigkeitswechsel 1967 (wechsl67) | 0 = nein, 1 = ja | ? | erhöht | 0,04 (0,05) | 0,19 (0,22) |
| Ereignisse während des Berufsverlaufs | Wehrdienst vor Tätigkeitswechsel (bundwehr) | 0 = nein, 1 = ja | kein Effekt | erhöht | 0,05 (0,07) | 0,21 (0,26) |
| | Weiterbildung vor Tätigkeitswechsel (weitbild) | 0 = nein, 1 = ja | erhöht | verringert | 0,04 (0,06) | 0,19 (0,23) |
| zeitabhängige persönliche Merkmale | Tätigkeitsdauer (tjob) | Dauer der Tätigkeit in Jahren | erhöht | kein Effekt | 5,9 (7,8) | 3,5 (3,1) |
| | Berufserfahrung Beginn (tlabor) | Dauer der Berufstätigkeit insgesamt in Jahren zu Beginn der Tätigkeit | erhöht | kein Effekt | 2,5 (0,0) | 3,4 (0,0) |
| Hilfsvariablen (s. Abschnitt 3.3) | Beginn (ts) | Beginn der Tätigkeit in Jahren seit Berufseintritt | – | – | 2,5 (0,0) | 3,4 (0,0) |
| | Ende (tf) | Ende der Tätigkeit in Jahren seit Berufseintritt | – | – | 8,5 (7,8) | 2,7 (3,1) |
| | Wechsel (sf1) | 0 = kein Wechsel, 1 = Abstieg, 2 = Aufstieg, 3 = horizontale Mobilität | – | – | – | – |

Mittelwert und Standardabweichung für erste Tätigkeiten in Klammern

ähnlicher Weise wird die Ableistung des Wehrdienstes gemessen. Da der Wehrdienst bis auf wenige Ausnahmen die Qualifikation einer Person nicht erhöht, erwarte ich keinen positiven Effekt auf die Aufstiegsrate. Dagegen ist anzunehmen, daß die Rückkehr in das Erwerbsleben nach Ableistung des Wehrdienstes mit Anpassungsproblemen verbunden ist und horizontale Mobilität oder sogar soziale Abstiege vermehrt auftreten können. MÜLLER (1979) hat z.B. festgestellt, daß die Ableistung des Wehrdienstes häufig mit einer späteren Dequalifikation bei Rückkehr in das Erwerbsleben verbunden ist.

Zusammengefaßt sind in diesem Anwendungsbeispiel alle Hypothesen vertreten, die ich in allgemeiner Form schon in der Einführung diskutiert habe.

a) *Zustandsabhängiger Prozeß*: Alle im Untersuchungszeitraum beobachteten Tätigkeitswechsel (Zustandswechsel) werden wie beschrieben in Auf- und Abstiege klassifiziert. Dabei wird angenommen, daß die Zustandswechsel, die mit einem Abstieg verbunden sind, mit geringerer Rate auftreten als Aufstiege und darüber hinaus auch in anderer Weise von den vorliegenden Kovariaten beeinflußt werden.

b) *heterogener Prozeß*: In der Untersuchung wird eine Fülle persönlicher, tätigkeits- und wirtschaftszweigspezifischer Einflußfaktoren kontrolliert, die Aufstiegs- und Abstiegsrate auf unterschiedliche Art und Weise beeinflussen. Einige dieser Merkmale sind konstant, während andere sich im Zeitablauf ändern.

c) *zeitabhängiger Prozeß*: Wie zu Anfang erläutert, nehmen Mobilitäts- und Aufstiegsrate bis zum Erreichen einer stabilen Erwerbsposition zu, während das Abstiegsrisiko konstant bleibt. Zumindest bei der Aufwärtsmobilität wird also ein zeitabhängiger Prozeß unterstellt, der, wie der folgende Punkt zeigt, in eine historische und eine aktuelle Komponente zerlegt werden kann.

d) *Vorgeschichte*: Durch Berücksichtigung der Abfolge der Tätigkeiten wird die Anzahl vorheriger Wechsel kontrolliert. Dabei wird angenommen, daß der Statusgewinn bei allen auf den ersten Tätigkeitswechsel folgenden Veränderungen geringer ist. Vorausgesetzt man hat die anfängliche Unterbewertung mit dem ersten Tätigkeitswechsel "wettgemacht", dann sind die folgenden Statusgewinne eher durchschnittlich. Wenn man darüber hinaus die gesamte Dauer der Berufstätigkeit (seit Berufseintritt) differenziert in Dauer der aktuellen Tätigkeit und Berufserfahrung zu Beginn dieser Tätigkeit, dann läßt sich der

o.g. zeitabhängige Prozeß in eine Vorgeschichte (Berufserfahrung) und eine aktuelle Komponente (Tätigkeitsdauer) zerlegen. Um beide Faktoren (Häufigkeit und akkumulierte Dauer vorheriger Zustände) untersuchen zu können, ist es jedoch notwendig, nicht nur die 7587 ersten sondern alle 13075 Tätigkeiten auszuwerten.

Abschließend sei noch einmal das Hauptargument der Untersuchung wiederholt: Danach hängen soziale Abstiege sehr viel mehr von externen, insbesondere wirtschaftlichen Merkmalen ab als von persönlichen Variablen. Dieser Umstand macht sich auf zweierlei Weise geltend: Einmal tragen Variablen wie Beschäftigtenwachstum oder Wechsel im Rezessionsjahr sehr viel mehr als alle anderen Variablen zur Erklärung sozialer Abstiege bei. Zum anderen treten solche externen Einflüsse in Prosperitätsphasen mehr oder weniger zufällig auf, so daß das Abstiegsrisiko im Gegensatz zu allen anderen Mobilitätsformen im Zeitablauf konstant bleibt.

## *3.3 Datenmanagement von Verlaufsdaten*

Nachdem ich in den beiden vorhergehenden Abschnitten die Daten der folgenden Auswertung etwas inhaltlicher dargestellt habe, möchte ich in diesem Kapitel wieder allgemeinere Probleme diskutieren. Dabei geht es um die Frage, wie man Verlaufsdaten aufbereitet, um sie mit entsprechenden Programmen auszuwerten. Anknüpfend an die typischen Fragestellungen aus Abschnitt 1.1.3 will ich zeigen, wie man die Dateien A−C am besten kodiert, um eine möglichst große Flexibilität der Datenanalyse zu gewährleisten. Dabei gehe ich zunächst davon aus, daß zeitkontinuierliche Verlaufsdaten vorliegen. Wie dieses Kodierungsschema für zeitdiskrete Daten zu modifizieren ist, wird in einem gesonderten Abschnitt 3.3.2 demonstriert.

Das Grundprinzip zeitkontinuierlicher und zeitdiskreter Datenstrukturen habe ich schon in der Einführung dargestellt (vgl. Tabelle 1.3). An dieser Stelle sollen vor allem zwei Fragen vertieft werden: 1) Wie ist diese Grundstruktur zu verallgemeinern, um möglichst viele unterschiedliche Analysestrategien zu erlauben? 2) Wie lassen sich die verschiedenen Erklärungsfaktoren, die man untersuchen möchte, in diese Datenstruktur einpassen? Hier ergeben sich vor allem bei den zeitabhängigen Kovariaten Probleme, auf die ich in Abschnitt 3.3.3 gesondert eingehen möchte. Abschnitt 3.3.4 betrachtet schließlich die Frage, welche zusätzlichen Daten-

managementprobleme auftreten, wenn sich der untersuchte Zustandsraum ändert oder wenn multivariate Verlaufsprozesse (multidimensionale Zustandsräume) analysiert werden.

### *3.3.1 Kodierung zeitkontinuierlicher Verlaufsdaten*

Wie aus Tabelle 1.3 zu entnehmen ist, bestehen die einzelnen Sätze einer *zeitkontinuierlichen Verlaufsdatei* aus den verschiedenen Zuständen, die die Untersuchungseinheiten im Zeitablauf einnehmen. Wie in Abschnitt 1.1.3 diskutiert, möchte man jedoch im Verlaufe einer Datenanalyse das Datenmaterial nach unterschiedlichen Fragestellungen *selektieren und aggregieren*. Es stellt sich also die Frage, welche notwendigen Informationen die einzelnen Datensätze enthalten müssen, um einerseits den Untersuchungsgegenstand adäquat zu erfassen und andererseits unterschiedliche Auswertungsstrategien zu ermöglichen.

Beginnen wir also mit der einfachsten Fragestellung nach der Dauer der ersten Tätigkeiten – technisch gesprochen also ein Prozeß mit *singulären, nicht–wiederholbaren* Ereignissen. Die entsprechenden Berufsverläufe von zwei zufällig ausgesuchten Personen der MZ71–Daten sind in Abbildung 3.2a dargestellt. Person 1 wechselt ihre Tätigkeit zum Zeitpunkt $t_1$, während Person 2 während des gesamten Untersuchungszeitraums von $t_0$ bis $\tau$ ihre zuerst eingenommene Tätigkeit nicht aufgibt. Diese (ersten) Tätigkeiten ergeben nun die Datensätze der Datei A. Zu ihrer vollständigen Charakterisierung benötigt man folgende Variablen: Anfangszeitpunkt TS, Endzeitpunkt TF, Anfangszustand SS, Endzustand SF. Die Dauer dieser Tätigkeiten (TF–TS) ist praktisch das abhängige Merkmal aller folgenden Auswertungen.

Tabelle 3.4 zeigt einen Ausschnitt der Datei A. Jede Person ist mit einer Tätigkeit (der ersten) vertreten. Der Endzeitpunkt TF entspricht jeweils dem Zeitpunkt des Wechsels (z.B. $t_1$) bzw. dem Zeitpunkt der letzten Beobachtung $\tau$ (= Ende des Untersuchungszeitraums). Dementsprechend zeigt der Endzustand SF, ob ein Ereignis stattgefunden hat (SF=1: Tätigkeitswechsel) oder nicht (SF=0: kein Wechsel). Da Anfangszeitpunkt TS und Anfangszustand SS für alle Untersuchungseinheiten identisch sind (TS=$t_0$, SS=0: kein Wechsel), hätte man diese Variablen auch aus der Datei weglassen können.

*Abbildung 3.2: Singuläre, multiple und wiederholbare Ereignisse*

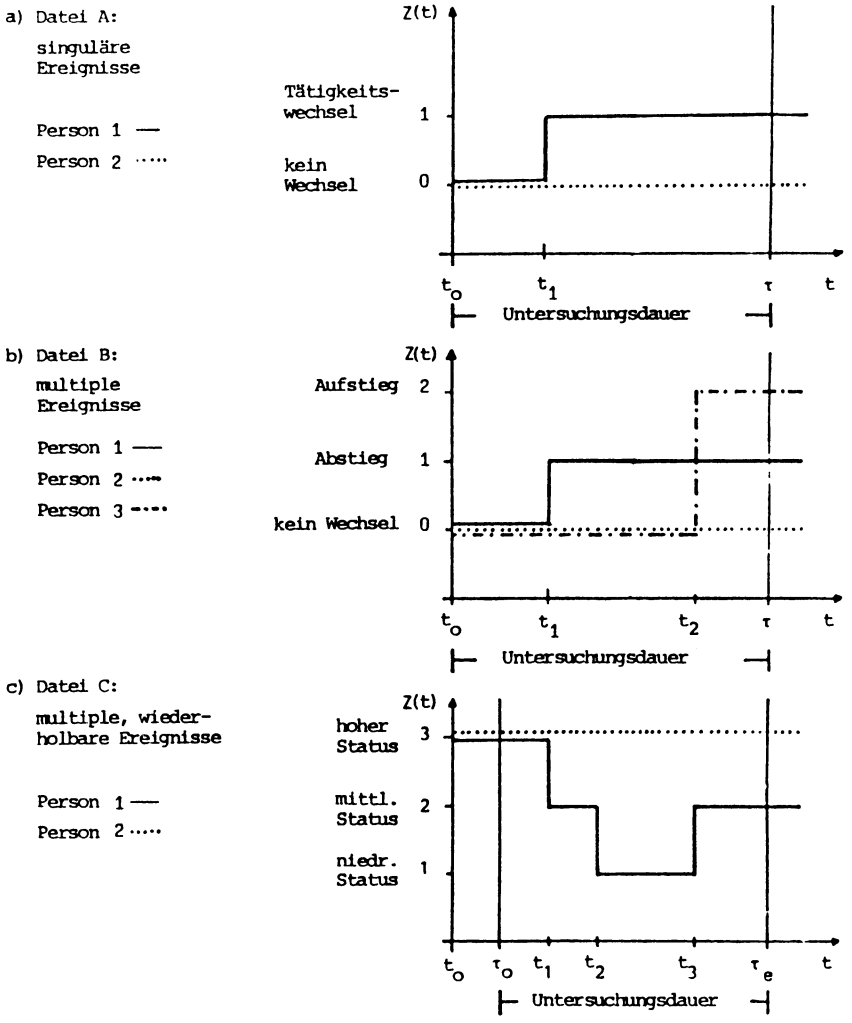


*Tabelle 3.4: Datei A – zeitkontinuierliche Verlaufsdaten*

| Person | Tätigkeit | TS | TF | SS | SF | $x_{i1}$ | $x_{i2}(TS)$ | $x_{i3}(TF)$ ... |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | $t_0$ | $t_1$ | 0 | 1 | $x_{11}$ | $x_{12}(t_0)$ | $x_{13}(t_1)$ ... |
| 2 | 1 | $t_0$ | $\tau$ | 0 | 0 | $x_{21}$ | $x_{22}(t_0)$ | $x_{23}(\tau)$ ... |
| . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . |

Tabelle 3.4 zeigt auch gleich, wie *erklärende Merkmale* **x** in der Datei zu kodieren sind. Sie werden jeweils an die schon vorhandenen Angaben zur Charakterisierung des abhängigen Merkmals angehängt. Das ist bei Variablen, die sich im Zeitablauf nicht verändern (vgl. $x_1$), kein Problem. Wenn man sich aber die Variablen der MZ71–Daten anschaut, sieht man jedoch, daß die wenigsten Merkmale wirklich zeitkonstant sind. Soziale Herkunft wäre allerhöchstens ein Beispiel. Die meisten Merkmale verändern sich im Zeitablauf, können also im Verlaufe einer Tätigkeit sehr unterschiedliche Werte annehmen, so daß das Problem entsteht, welchen dieser Werte man für die folgende Auswertung verwendet. Richtig wäre es natürlich, den gesamten Verlauf dieser zeitveränderlichen Merkmale zu berücksichtigen, aus praktischen Gründen verwendet man jedoch häufig *ad–hoc–Lösungen*, die nur einen "repräsentativen" Wert berücksichtigen – z.B. den Wert zu Beginn oder am Ende eines Zustands bzw. den Durchschnitt aller Werte während eines Zustands. Dieses Vorgehen ist unter bestimmten Bedingungen problematisch und ich werde darauf in Abschnitt 3.3.3 zurückkommen.

Eine Analyse verschiedener Arten von (ersten) Tätigkeitswechseln (*multiple Ereignisse*) läßt sich ebenfalls relativ einfach mit der oben beschriebenen Datenstruktur umsetzen. Abbildung 3.2b zeigt dazu die Berufsverläufe von drei Personen, die ihre erste Tätigkeit im Untersuchungszeitraum mit einem Aufstieg oder Abstieg bzw. ohne Wechsel beenden. Tabelle 3.5 zeigt die entsprechende Datei B. Es hat sich im wesentlichen nichts verändert, außer daß der Endzustand jetzt mehrere Ausprägungen hat. Auch hier hätte man die Variablen TS und SS weglassen können, denn diese Informationen sind redundant. Man beachte allerdings, daß Datei A jederzeit aus Datei B generiert werden kann, wenn man eine Variable SF1 einführt, die die verschiedenen Arten von Tätigkeitswechseln SF zusammenfaßt. Bei Auswertung A verwendet man die abhängigen Merkmale TF und SF1, bei Auswertung B hingegen TF und SF.

*Tabelle 3.5: Datei B – zeitkontinuierliche Verlaufsdaten*

| Person | Tätigkeit | TS | TF | SS | SF | SF1 | $x_{i1}$ | $x_{i2}(TS)$ | $x_{i3}(TF)$ ... |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | $t_0$ | $t_1$ | 0 | 1 | 1 | $x_{11}$ | $x_{12}(t_0)$ | $x_{13}(t_1)$ ... |
| 2 | 1 | $t_0$ | $\tau$ | 0 | 0 | 0 | $x_{21}$ | $x_{22}(t_0)$ | $x_{23}(\tau)$ ... |
| 3 | 1 | $t_0$ | $t_2$ | 0 | 2 | 1 | $x_{31}$ | $x_{32}(t_0)$ | $x_{33}(t_2)$ ... |
| . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . |

Abbildung 3.2c berücksichtigt schließlich alle Komplikationen einer Verlaufsanalyse: Es sind *multiple Ereignisse* (Wechsel zwischen verschiedenen Statuspositionen) möglich, die sich überdies im Zeitablauf *wiederholen* können. Außerdem wird deutlich, daß der gesamte Prozeß nur begrenzt erhoben wurde: Es gibt links– bzw. rechtszensierte Beobachtungen. *Rechtszensierte Beobachtungen* lassen sich einfach berücksichtigen, indem man den Endzustand SF mit dem Anfangszustand SS gleichsetzt. *Linkszensierte Beobachtungen* lassen sich hingegen nicht so einfach behandeln. Man macht daher in der Regel die vereinfachende Annahme, daß der Beginn des Prozesses $t_0$ mit dem Untersuchungsbeginn $\tau_0$ zusammenfällt. Davon unabhängig lassen sich auch diese sehr viel komplexeren Daten in die mittlerweile bekannte Datenstruktur einpassen. Tabelle 3.6 zeigt diese Datei C und wie gleich deutlich wird, sind auch in dieser Datei praktisch alle vorhergehenden Analysen enthalten.

Zunächst möchte ich jedoch auf die Spezifika von Datei C eingehen. Da jetzt alle Ereignisse im Untersuchungszeitraum berücksichtigt werden, sind pro Untersuchungseinheit eine unterschiedliche Anzahl von Datensätzen vertreten, je nachdem, wie hoch die Mobilitätsrate der jeweiligen Person ist. Anfangszeitpunkt TS und Anfangszustand SS spielen jetzt erstmals eine Rolle. Aus der Kombination von Anfangs– und Endzustand ergeben sich alle Übergänge, die in dem Beispiel möglich sind, so daß jeweils die richtigen Daten zur Schätzung der 6 verschiedenen Übergangsraten bereitgestellt werden können. Dabei kann allerdings das abhängige Merkmal Dauer auf unterschiedliche Art und Weise berechnet werden. Einmal kann es sich um die gesamte Prozeßdauer TF bis zum Eintritt des jeweiligen Ereignisses handeln, zum anderen kann es sich auch um die Dauer TF1 des aktuellen Zustands handeln. TF entspräche in unserem Fall der Berufserfahrung, TF1 der Tätigkeitsdauer. Will man beide Einflüsse gleichzeitig über-

prüfen, dann ist es sinnvoll, die Zustandsdauer TF1 von der Prozeßdauer TF abzuziehen, da TF1 in TF enthalten ist. Zusammengefaßt lassen sich Veränderungen der Mobilitätsrate mit der Tätigkeitsdauer und/ oder der Berufserfahrung untersuchen, wobei man mit Hilfe von SS und SF zwischen verschiedenen Tätigkeitswechseln unterscheiden kann.

*Tabelle 3.6: Datei C — zeitkontinuierliche Verlaufsdaten*

| Person | Tätigkeit | TS | TF | TF1 | SS | SS1 | SF | SF1 | SF2 | $x_{i1}$ | $x_{i2}(TS)$ | $x_{i3}(TF)$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | $\tau_0$ | $t_1$ | $t_1-\tau_0$ | 3 | 0 | 2 | 1 | 1 | $x_{11}$ | $x_{12}(\tau_0)$ | $x_{13}(t_1)\ldots$ |
| 1 | 2 | $t_1$ | $t_2$ | $t_2-t_1$ | 2 | 0 | 1 | 1 | 1 | $x_{11}$ | $x_{12}(t_1)$ | $x_{13}(t_2)\ldots$ |
| 1 | 3 | $t_2$ | $t_3$ | $t_3-t_2$ | 1 | 0 | 2 | 1 | 2 | $x_{11}$ | $x_{12}(t_2)$ | $x_{13}(t_3)\ldots$ |
| 1 | 4 | $t_3$ | $\tau_e$ | $t_4-\tau_e$ | 2 | 0 | 2 | 0 | 0 | $x_{11}$ | $x_{12}(t_3)$ | $x_{13}(\tau_e)\ldots$ |
| 2 | 1 | $\tau_0$ | $\tau_e$ | $\tau_e-\tau_0$ | 3 | 0 | 3 | 0 | 0 | $x_{21}$ | $x_{22}(\tau_0)$ | $x_{23}(\tau_e)\ldots$ |
| . | . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . | . | . |

Natürlich sind auch alle vorhergehenden Auswertungen in Datei C enthalten. Dort wurden jeweils nur die ersten Tätigkeiten berücksichtigt. Die gleichen Datensätze lassen sich jetzt über die Variable Tätigkeit = 1 auswählen. Auswertung A (Dauer der ersten Tätigkeiten) ergibt sich dann, wenn man für alle verbliebenen Datensätze den gleichen Anfangszustand SS1 = 0 verwendet und beim Endzustand SF1 lediglich vermerkt, ob eine Veränderung stattgefunden hat (SF1 = 1) oder nicht (SF1 = 0). Die Dauer der ersten Tätigkeiten entspricht TF1. Für Auswertung B (Auf— und Abstiege bei ersten Tätigkeiten) verwendet man die gleichen Angaben, lediglich der Endzustand SF2 wird neu definiert. Dabei wird aus der Kombination von SS und SF berechnet, ob es sich um einen Abstieg (SF2 = 1), einen Aufstieg (SF2 = 2) oder gar keinen Wechsel handelt (SF2 = 0).

Zusammengefaßt lassen sich also die verschiedensten Auswertungen mit Datei C durchführen, ohne daß die Daten abgeändert werden müssen. Voraussetzung ist allerdings, daß die verschiedenen Analysen nur eine Aggregation und Disaggregation der Zustandsvariablen SS und SF erfor-

dern.[1] Das ist häufig nicht der Fall. In vielen Fällen will man mit den gleichen Daten andere Zustandsräume untersuchen, die Anzahl und Zeitpunkte der Ereignisse im Untersuchungszeitraum völlig neu definieren. Ein Beispiel wäre etwa eine Untersuchung des Arbeitslosigkeitsrisikos mit den MZ71 – Daten. Man würde dann nur noch Beschäftigungs – und Arbeitslosigkeitszeiten unterscheiden. Mehrere aufeinanderfolgende Tätigkeiten ergäben eine Beschäftigungszeit, sofern sie nicht durch Zeiten der Arbeitslosigkeit unterbrochen wird. Mit dieser *Neudefinition des Zustandsraumes* wären die ursprünglichen Daten wertlos, die auf der Basis von Tätigkeiten erstellt wurden. Es ist eine Reorganisation des Datenfiles notwendig (vgl. Abschnitt 3.3.4).

### *3.3.2 Kodierung zeitdiskreter Verlaufsdaten*

Im Gegensatz zu zeitkontinuierlichen Verlaufsdaten mit einem Datensatz pro Ereignis sind *zeitdiskrete Verlaufsdaten* nach Erhebungszeitpunkten organisiert (vgl. Tabelle 1.3). Ich will diese Situation noch einmal an Hand der MZ71 – Daten verdeutlichen. Die Erhebung der einzelnen Tätigkeiten erfolgte bekanntlich jahresweise. Von daher bietet sich eine zeitdiskrete Kodierung förmlich an. Die drei Veränderungen, die Person 1 in Abbildung 3.2c erfährt, seien im 2., 5. bzw. 8. Jahr eingetreten. Jeder Datensatz dieser zeitdiskreten Verlaufsdatei D gibt nun an, welche Tätigkeit in dem jeweiligen Intervall oder zu dem jeweiligen Stichtag ausgeübt wurde (vgl. Tabelle 3.7). Um Veränderungen festzustellen, ist es notwendig, den aktuellen Zustand SS der Untersuchungseinheit mit dem der Vorperiode LAST_SS zu vergleichen. Je nachdem ergibt sich keine Veränderung (SF1 = 0), ein Abstieg (SF1 = 1) oder ein Aufstieg (SF1 = 2).

Auch hier kann man zeigen, daß Dateien geringerer Komplexität jederzeit aus Datei D generiert werden können. Betrachten wir jedoch die besonderen Eigenschaften dieser zeitdiskreten Verlaufsdatei: Als erstes fällt dabei auf, daß sich die Fallzahl beim Übergang auf einzelne Erhebungszeitpunkte (statt Ereignisse bei zeitkontinuierlichen Verlaufsdaten) wesentlich erhöht. Pro Untersuchungseinheit tritt immer die gleiche Anzahl von Datensätzen auf, die genau der Anzahl K der Erhebungszeitpunkte entspricht.

---

1) Es ist natürlich nicht besonders effizient, die Variablen SS1, SF1, SF2 extra zu kodieren, da man diese Merkmale auch durch entsprechende Befehle aus SS und SF berechnen kann.

Die ursprüngliche Datenbasis vom Umfang N vervielfacht sich also um den Faktor K.

*Tabelle 3.7: Datei D – zeitdiskrete Verlaufsdaten*

| Person | Erhebung | LAST_SS | SS | SF1 | $x_{i1}$ | $x_{i2}(t)$ | $x_{i3}(t)$ |
|---|---|---|---|---|---|---|---|
| 1 | 1 | . | 3 | 0 | $x_{11}$ | $x_{12}(1)$ | $x_{13}(1)$ |
| 1 | 2 | 3 | 2 | 1 | $x_{11}$ | $x_{12}(2)$ | $x_{13}(2)$ |
| 1 | 3 | 2 | 2 | 0 | $x_{11}$ | $x_{12}(3)$ | $x_{13}(3)$ |
| 1 | 4 | 2 | 2 | 0 | $x_{11}$ | $x_{12}(4)$ | $x_{13}(4)$ |
| 1 | 5 | 2 | 1 | 1 | $x_{11}$ | $x_{12}(5)$ | $x_{13}(5)$ |
| 1 | 6 | 1 | 1 | 0 | $x_{11}$ | $x_{12}(6)$ | $x_{13}(6)$ |
| 1 | 7 | 1 | 1 | 0 | $x_{11}$ | $x_{12}(7)$ | $x_{13}(7)$ |
| 1 | 8 | 1 | 2 | 2 | $x_{11}$ | $x_{12}(8)$ | $x_{13}(8)$ |
| 1 | 9 | 2 | 2 | 0 | $x_{11}$ | $x_{12}(9)$ | $x_{13}(9)$ |
| 1 | 10 | 2 | 2 | 0 | $x_{11}$ | $x_{12}(10)$ | $x_{13}(10)$ |
| 2 | 1 | . | 3 | 0 | $x_{21}$ | $x_{22}(1)$ | $x_{23}(1)$ |
| 2 | 2 | 3 | 3 | 0 | $x_{21}$ | $x_{22}(2)$ | $x_{23}(2)$ |
| 2 | 3 | 3 | 3 | 0 | $x_{21}$ | $x_{22}(3)$ | $x_{23}(3)$ |
| . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . |
| 2 | 10 | 3 | 3 | 0 | $x_{21}$ | $x_{22}(10)$ | $x_{23}(10)$ |
| . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . |

Zur Beschreibung des abhängigen Merkmals braucht man eigentlich nur die Zustandsvariable SS, die durch Vergleich mit dem Wert der Vorperiode LAST_SS Auskunft über die Art des Ereignisses gibt.[1] Eine explizite Angabe des Veränderungszeitpunktes wie bei zeitkontinuierlichen Verlaufsdaten ist nicht notwendig, da dieser sich implizit aus der Abfolge der Erhebungen ergibt. Natürlich ist es aus Gründen der Lesbarkeit sinnvoller, statt der Nummer der Erhebung den jeweiligen Erhebungszeitpunkt zu kodieren. In Tabelle 3.7 stimmen beide Angaben zufällig überein.

1) Kenner der Zeitreihenanalyse wissen, daß man die Variable LAST_SS nicht explizit kodieren muß. Sie läßt sich durch einen entsprechenden LAG – Befehl aus SS berechnen. Das gleiche gilt im übrigen auch für die Variable SF1, die sich aus dem Vergleich von SS und LAST_SS ergibt.

An dieser Stelle liegt es nahe, die Beschreibung des abhängigen Merkmals in beiden Dateien zu vergleichen (vgl. Tabelle 3.6 und 3.7). SS und LAST_SS in D entsprechen offenbar den Angaben über End- und Anfangszustand SF und SS bei zeitkontinuierlichen Verlaufsdaten. Für die Dauer der jeweiligen Zustände (TF-TS) gibt es bei zeitdiskreten Verlaufsdaten kein explizites Äquivalent. Wie man sich jedoch leicht überlegen kann, entspricht diese Dauer dem zeitlichen Abstand der einzelnen Erhebungen. Man kann die zeitkontinuierliche Datei C sogar in die zeitdiskrete Datei D überführen, indem man die ursprünglich mehr-jährigen Tätigkeiten in mehrere kurze Abschnitte von einem Jahr aufteilt.

Bei dem Vergleich der *erklärenden Merkmale* in Tabelle 3.6 und 3.7 fällt eine weitere Besonderheit zeitdiskreter Verlaufsdaten auf. Da hier viele Informationen pro Untersuchungseinheit zur Verfügung stehen (pro Erhebungszeitpunkt ein Datensatz, statt weniger Zustände, die jeweils einen längeren Zeitraum abdecken), kann man auch mehrere Werte bei den zeitabhängigen erklärenden Variablen berücksichtigen. Die Kodierung der zeitabhängigen Merkmale ist damit vom Erhebungsplan und nicht von den Ereignissen des untersuchten Veränderungsprozesses abhängig.

Schließlich läßt sich zeigen, daß *Neudefinitionen des untersuchten Zustandsraumes* keine so drastische Reorganisation des Datenfiles erfordern wie bei zeitkontinuierlichen Verlaufsdaten. Voraussetzung ist allerdings, daß entweder die Erhebungsintervalle hinreichend klein sind oder die Neudefinition des Zustandsraumes nur eine Aggregierung der ursprünglichen Zustandsdauern erfordert. Das wäre beispielsweise der Fall, wenn man von einer Analyse der Tätigkeiten (mit Auf- und Abstiegen) zu einer Analyse von Beschäftigung und Arbeitslosigkeit übergehen würde. Dazu ist es bei der Datenaufnahme nötig, alle Tätigkeiten (inkl. Arbeitslosigkeit und Nicht-Erwerbstätigkeit) so differenziert wie möglich zu erfassen. Die Variable SS hat also zu Beginn sehr viele Ausprägungen und durch einen entsprechenden LAG-Befehl kann man jeden aktuellen Wert SS mit dem vorhergehenden LAST_SS vergleichen. Für die erste der beiden Analysen würde man nur die Veränderungen als Ereignis werten, die einen Aufstieg bzw. einen Abstieg zur Folge haben, und dementsprechend die Variable SF1 bilden. Für die zweite Analyse hingegen tritt immer dann ein Ereignis auf, wenn ein Übergang zwischen einer Erwerbstätigkeit (unabhängig von der konkreten Tätigkeit) und einer Arbeitslosigkeit stattfindet. Da häufig mehrere Tätigkeiten aufeinanderfolgen, ohne daß eine Person arbeitslos wird, ist keine Reorganisation der Daten notwendig. Die neu zu bildende Variable SF2 enthält entsprechend weniger Ereignisse.

Der Preis dieser flexiblen Dateiorganisation bzw. der adäquaten Behandlung zeitabhängiger Variablen ist jedoch eine immense Steigerung der Fallzahl, insbesondere wenn man berücksichtigt, daß die Erhebungsintervalle eigentlich so klein sein sollten, daß sie jede theoretisch interessante Veränderung erfassen. Doch kommt diese Situation in der Forschungspraxis nicht so häufig vor – glücklicherweise könnte man fast sagen. Man überlege sich jedoch einmal, welche Konsequenzen diese Strategie für den oben beschriebenen Datensatz hätte: Würde man die Meßfehlerproblematik ernst nehmen und die Daten als zeitdiskrete Verlaufsdaten behandeln, ergäben sich ca. 10867 Personen · 10 Erhebungszeitpunkte = 108670 Datensätze. Eine solche Datei würde die EDV–Ressourcen (Speicherplatz, Rechenzeit) über alle Maßen beanspruchen. Eine effiziente Speicherung ist nur dann möglich, wenn alle (unabhängigen und abhängigen) Variablen diskret sind. Dann ergibt sich eine *multivariate Kreuztabelle*, wie sie Tabelle 3.8 zeigt. Hier hat das erklärende Merkmal "Ausbildung" nur zwei Ausprägungen. Kombinationen mit anderen diskreten Merkmalen sind denkbar, ändern jedoch nichts an der grundlegenden Struktur, sondern führen lediglich zu einer weiteren Auffächerung der Tabelle.

*Tabelle 3.8: Dauer erster Tätigkeiten und Art des Wechsels von Personen mit und ohne abgeschlossene Berufsausbildung*

| | ohne BA | | mit BA | | Insgesamt | |
|---|---|---|---|---|---|---|
| Jahre | Abstieg | Sonstige[a)] | Abstieg | Sonstige[a)] | N | % |
| <1 | 0 | 0 | 0 | 0 | 0 | 0,0% |
| 1–2 | 19 | 35 | 110 | 264 | 428 | 4,0% |
| 2–3 | 21 | 39 | 105 | 357 | 522 | 4,9% |
| 3–4 | 35 | 44 | 135 | 398 | 612 | 5,8% |
| 4–5 | 23 | 21 | 91 | 413 | 548 | 5,2% |
| 5–6 | 12 | 35 | 107 | 394 | 548 | 5,2% |
| 6–7 | 24 | 34 | 75 | 411 | 544 | 5,1% |
| 7–8 | 26 | 34 | 99 | 448 | 607 | 5,7% |
| 8–9 | 25 | 53 | 118 | 475 | 671 | 6,3% |
| 9–10 | 24 | 201 | 91 | 1779 | 2095 | 19,8% |
| 10–11 | 13 | 178 | 52 | 1728 | 1971 | 18,6% |
| >11 | 0 | 160 | 0 | 1869 | 2029 | 19,2% |
| Insgesamt | 222 | 834 | 983 | 8536 | 10575[b)] | 100,0% |

a) Aufstiege, horiz. Mobilität und ohne Wechsel. b) 91 Missings bei der Variablen "Qualifikation"

### *3.3.3 Kodierung zeitabhängiger Kovariaten*

Wie die Diskussion der beiden vorherigen Abschnitte zeigt, besteht die kleinste Einheit einer Verlaufsdatei zunächst einmal aus Angaben über den zeitlichen Verlauf des in Frage stehenden Merkmals. An diese Datensätze werden nun weitere Angaben über mögliche Erklärungsfaktoren angehängt. Daß diese *Kodierung zeitabhängiger Variablen* bei zeitkontinuierlichen Verlaufsdaten nicht ohne Probleme ist, wurde bereits angedeutet und soll nun vertieft werden.

Betrachten wir als Beispiel Einkommen: Es ist plausibel anzunehmen, daß das Einkommen mit zunehmender Berufserfahrung ebenfalls ansteigt. Je länger es also dauert, bis ein Tätigkeitswechsel eintritt, um so höher ist auch der persönliche Verdienst. Übernimmt man jetzt die in den Dateien A–C vorgenommenen *ad–hoc–Lösungen*, dann ist der konkrete Wert, der für x(t) kodiert wird, von den Ereignissen des Prozesses abhängig. Das ursprünglich exogen gedachte Merkmal wird durch das Kodierverfahren endogen. Wenn das Merkmal zudem noch einen bestimmten (positiven) Trend aufweist, wie z.B. Einkommen, dann ist es notwendigerweise so, daß immobile Personen (solche mit längerer Tätigkeitsdauer) höhere Einkommenswerte aufweisen. Obwohl die Mobilitätsrate durch das Einkommen erklärt werden soll, ist auf Grund des Kodierverfahrens das Einkommen aus der Mobilitätsrate ableitbar. Im Ergebnis wird der Einfluß des Einkommens verzerrt geschätzt (FLINN/HECKMANN 1982: 64ff.)

Das Argument gilt für alle der o.g. ad–hoc–Lösungen.[1] Wie läßt sich dieses Problem umgehen? Dazu gibt es drei *Alternativen*:

1. Man kann natürlich annehmen, daß das exogene Merkmal x(t) zwar im Zeitablauf variiert, dabei aber keinen besonderen Trend aufweist. Diese Annahme ist gleichbedeutend mit der Aussage, daß das Merkmal x(t) im Prinzip zeitkonstant ist, allerhöchstens zu verschiedenen Zeitpunkten auf Grund von Meßfehlern verschiedene Werte erhoben werden.
2. Man macht eine Annahme darüber, wie sich die Variable x(t) im Zeitablauf ändert. Man gibt also an, wie beispielsweise das Einkom-

1) Bei Kodierung des Anfangszustands: keine Probleme für die ersten Ereignisse, aber die Werte für alle folgenden Ereignisse hängen von der Dauer der vorherigen Zustände ab. Bei Kodierung des Durchschnitts aller Werte im Zeitraum(TS, TF): dieser Durchschnit ist von der Dauer (TF–TS) abhängig.

men mit der Berufserfahrung zunimmt. Nachdem diese funktionale Abhängigkeit spezifiziert ist, kann man diese Annahme in ein Regressionsmodell für Verlaufsdaten einfügen.

3. Häufig hat man jedoch über die Art der Veränderung der exogenen Merkmale keine Informationen. Man möchte also möglichst alle im Zeitablauf auftretenden Werte in die Berechnungen eingehen lassen. Wie das möglich ist, wird aus der Diskussion zeitdiskreter Verlaufsdaten deutlich. Dort existiert nämlich das Endogenitätsproblem nicht – und zwar deshalb, weil die Kodierung der exogenen Merkmale nicht von den Ereignissen des Prozesses sondern von extern gesetzten Erhebungszeitpunkten abhängig ist. Da sich jeder zeitkontinuierliche Verlaufsdatensatz in kleine zeitdiskrete Stücke zerteilen läßt, können auf diese Weise möglichst viele der im Zeitablauf auftretenden Werte von x(t) berücksichtigt werden.

Mit der letzten Alternative wird auch das allgemeine methodische Problem angedeutet, das hinter der praktischen Frage steht, wie man zeitabhängige exogene Merkmale kodiert. Dieses Problem läßt sich auf folgende eingängige Formel bringen: Zeitkontinuierliche Verlaufsdaten sind auf spezifische Weise sehr genau, aber andererseits auch wiederum sehr ungenau. Während man einerseits das abhängige Merkmal mit genauen Veränderungszeitpunkten erfaßt, sind andererseits die unabhängigen Merkmale und deren Veränderung sehr ungenau abgebildet. Richtig wäre vielmehr, Verlaufsdaten über das abhängige und die unabhängigen Merkmale zu erheben. Dabei entsteht jedoch das Problem, wie man die verschiedenen zeitlichen Verläufe zusammenführt (vgl. den folgenden Abschnitt). Umgekehrt wird deutlich, warum dieses Problem bei zeitdiskreten Verlaufsdaten nicht entstehen kann. Hier werden nämlich alle (abhängigen und unabhängigen) Merkmale dem gleichen (zeitdiskreten) Beobachtungsraster unterworfen.

### *3.3.4 Allgemeine Probleme des Datenmanagements von Verlaufsdaten*

Abschließend möchte ich noch einmal die wesentlichen Datenmanagementprobleme zusammenfassen:

1. Längsschnittdaten (nicht nur Verlaufsdaten) sind in der Regel sehr viel umfangreicher als übliche Querschnittsdaten, da sie zusätzlich zur Objektdimension noch die Zeitdimension erfassen. Eine Panelunter-

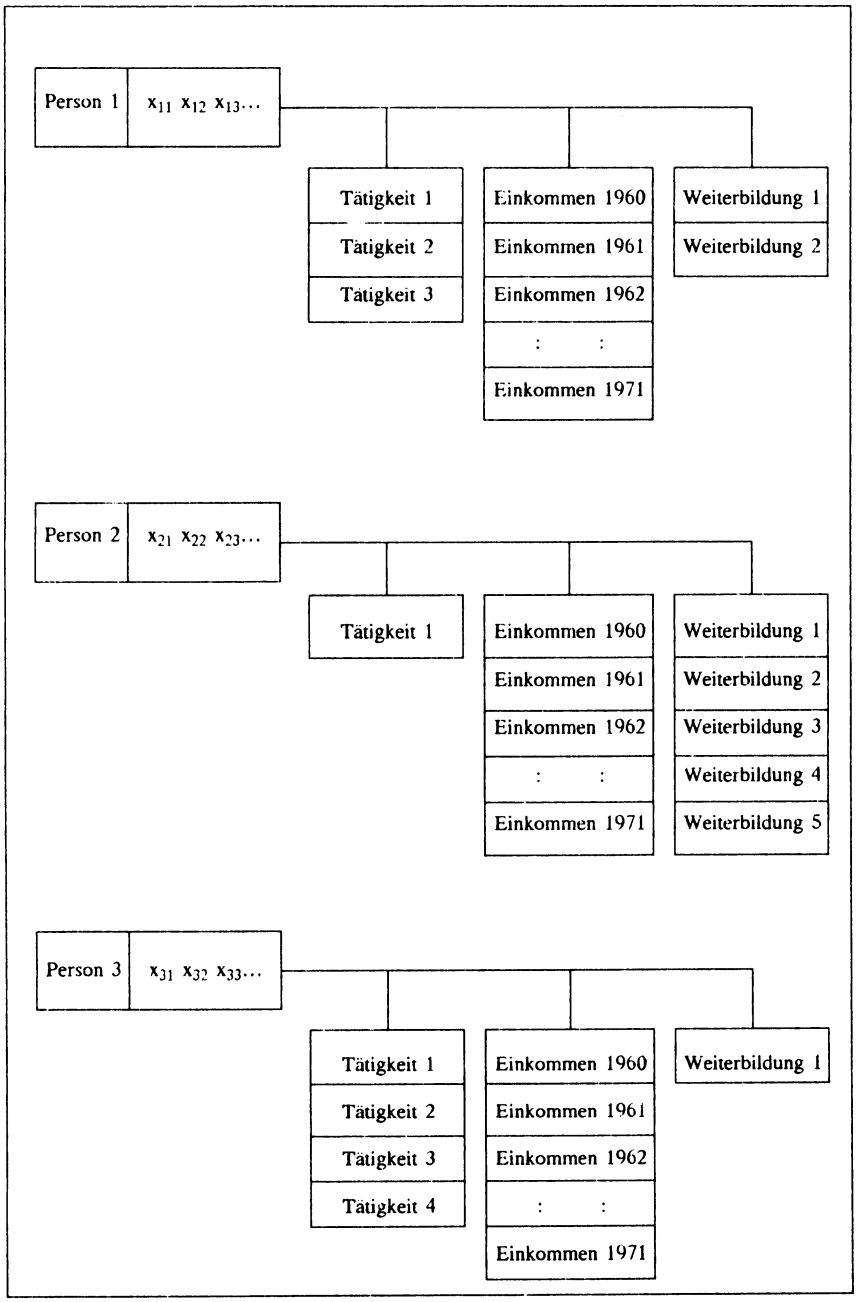
suchung mit 2 Befragungswellen beispielsweise produziert gleich doppelt so viele Daten wie eine entsprechende Querschnittsuntersuchung. Ohne eine effiziente Speicherung der Information wird man der Datenfülle nicht Herr.

2. Alle o.g. Dateien gehen darüber hinaus mit dem Speicherplatz sehr verschwenderisch um, indem sie entweder erklärende Merkmale auf jedem Datensatz wiederholen, obwohl sie sich im Zeitablauf nicht ändern, oder indem sie wie bei zeitdiskreten Daten eine unnötige Anzahl von Nicht–Ereignissen enthalten.
3. Je nach Definition des interessierenden Zustandsraumes sind möglicherweise Aggregationen oder Disaggregationen entlang der Zeitdimension notwendig. Bei den oben diskutierten Dateien hätte das zur Folge, daß man entweder ursprünglich getrennte Datensätze wieder zusammenfügen oder schon vorhandene Datensätze noch weiter zerteilen müßte. Es sei denn, man bevorzugt von Anfang an eine Datenaufnahme, die den gesamten Prozeß soweit zergliedert, daß auch jede noch so kleine Veränderung abgebildet wird. Damit wäre man jedoch wieder bei Punkt 2 angelangt.

Datentechnisch gesprochen handelt es sich bei Verlaufsdaten um hierarchische Datenfiles. Aus der folgenden Abbildung 3.3 werden die zwei wesentlichen Charakteristika einer solchen Datenstruktur sichtbar.

- Die einzelnen Personen sind immer noch die wesentlichen Analyseeinheiten eines Verlaufsdatensatzes. Sie können mit wenigen konstanten Merkmalen beschrieben werden. Ihnen zugeordnet sind eine Reihe weiterer Informationen, die den zeitlichen Verlauf des abhängigen Merkmals (hier: Tätigkeit) sowie der anderen zeitabhängigen erklärenden Merkmale beschreiben (hier: Zeitreihe Einkommen, Verlauf Weiterbildung). Die Analyseeinheiten sind dort die einzelnen Ereignisse (Tätigkeit, Weiterbildung) bzw. die Erhebungszeitpunkte (Einkommen). Eine solche Datenstruktur, in der Datensätze anderen logisch untergeordnet sind, nennt man *hierarchisch.*
- Die Zahl dieser logisch untergeordneten Informationen variiert von Person zu Person bzw. zwischen den einzelnen zeitbezogenen Merkmalen Tätigkeit, Einkommen und Weiterbildung. Der gesamte Datensatz einer Person (konstante und zeitbezogene Merkmale zusammen) hat daher *variable Satzlänge.*

*Abbildung 3.3: Verlaufsdaten als hierarchische Datenfiles*



Wenn man alle vier unterschiedlich häufig auftretenden Merkmale bzw. Merkmalsgruppen in jeweils einer Personendatei, einer Tätigkeitsdatei, einer Einkommensdatei und einer Weiterbildungsdatei ablegt, dann erzielt man eine komprimierte, aber dennoch vollständige Abspeicherung aller vorhandenen Informationen, ohne einzelne Werte unnötig zu wiederholen. Durch ein Datenbanksystem ist lediglich sicherzustellen, daß die Daten jederzeit und unter unterschiedlichen Fragestellungen zusammengeführt werden können, um – fertig aufbereitet – an ein Auswertungsprogramm zur weiteren Analyse übergeben werden zu können. Diese Aufbereitung führt das Datenbanksystem mit Hilfe von Querverweisen zwischen den Dateien durch. In diesem Beispiel bestünden diese sogenannten Pointer–Variablen aus einer Personenidentifikationsnummer und der Zeit. Ein solches Datenbanksystem für den sozialwissenschaftlichen Bereich wäre z.B. das Programmpaket SIR. Mit den mittlerweile eingetretenen Weiterentwicklungen bei den statistischen Programmpaketen läßt sich ein solches Datenbankkonzept aber auch mit SPSS und insbesondere mit SAS umsetzen. Im Rahmen einer solchen Datenbank ist es dann auch möglich, verschiedene Aggregationen und Disaggregationen entlang der Zeitdimension vorzunehmen (vgl. das Beispiel Arbeitslosigkeit und Beschäftigung).

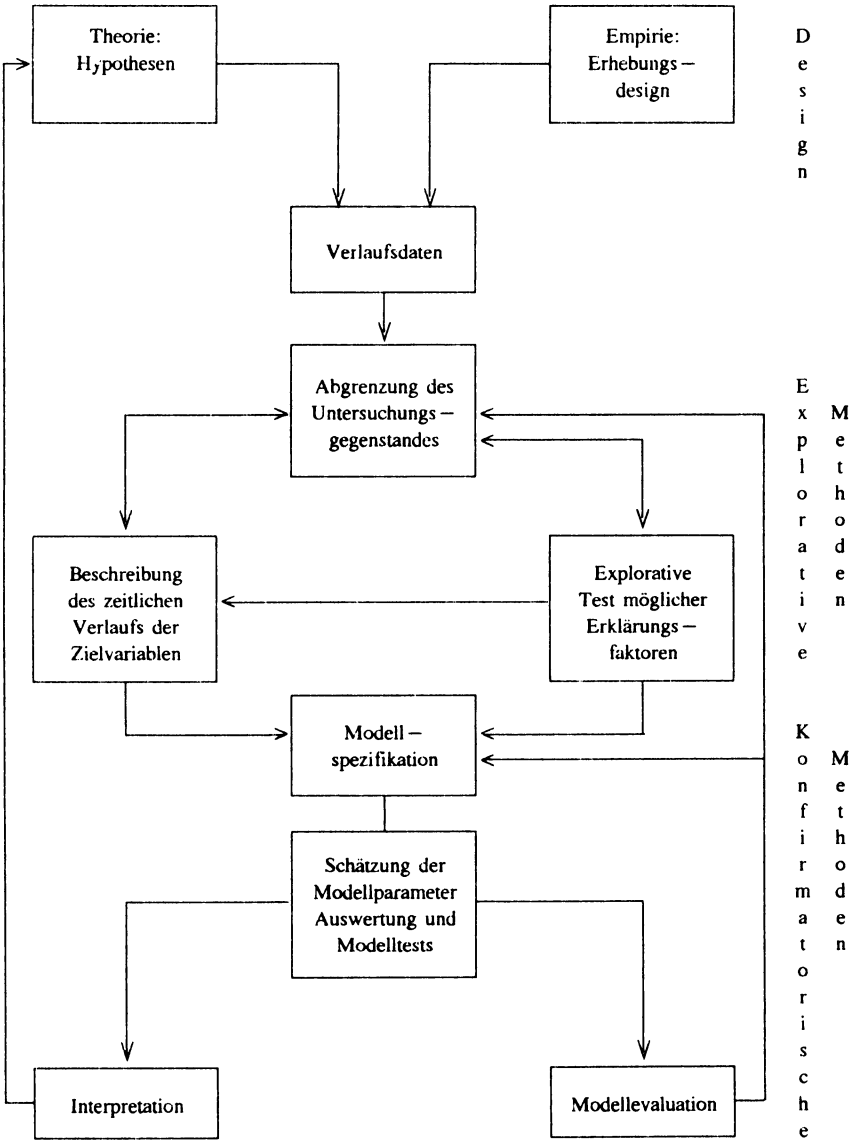
## *3.4 Statistische Auswertung von Verlaufsdaten – Vorgehen und verfügbare Programme*

Abbildung 3.4 zeigt den typischen Ablauf einer Analyse von Verlaufsdaten. Da diese Art von Längsschnittdaten in der Regel sehr viel mehr Informationen als übliche Querschnittsdaten enthalten, lassen sich eine Vielzahl unterschiedlicher Untersuchungsfragen bearbeiten. Ein Großteil der Auswertungsarbeit wird sich daher auf die *Abgrenzung des Untersuchungsgegenstandes* konzentrieren:

- Welche *Zielvariablen* erfassen also den interessierenden Gegenstandsbereich am besten?
- Wie läßt sich ihr *zeitlicher Verlauf* angemessen beschreiben?
- Welche möglichen *Erklärungsfaktoren* stehen zur Verfügung und zeigen sich bei einer ersten explorativen Datenanalyse die vermuteten Zusammenhänge?

Dieser Teil der Auswertung ist in der Regel ein iterativer Prozeß, in dem die ursprünglichen Untersuchungsfragen häufig abgeändert werden, weil sie

*Abbildung 3.4: Stadien einer Verlaufsanalyse*



sich empirisch nicht halten lassen oder weil neue Erkenntnisse andere Herangehensweisen an das Datenmaterial nahelegen. Es versteht sich von selbst, daß alle drei Elemente dieser Auswertung in einem gegenseitigen Abhängigkeitsverhältnis zueinander stehen: Eine neu definierte Zielvariable setzt z.B. andere Erklärungsfaktoren voraus und umgekehrt. Am Ende dieses *explorativen Teils* einer Verlaufsdatenanalyse stehen in der Regel präzisierte Untersuchungsfragen *(Modelle)*, die nun im nächsten Schritt mit den klassischen Mitteln der *hypothesentestenden* Statistik überprüft werden sollen. Meine These ist aber, daß der explorative Teil zeitlich gesehen den weitaus größten Teil der Auswertung ausmacht.

Ganz grob lassen sich also Methoden der Verlaufsdatenanalyse in mehr *explorative* und mehr *konfirmatorische* (hypothesentestende) Verfahren unterteilen. Im Vorgriff auf die folgenden zwei Kapitel ergeben sich dabei vier Aufgabenbereiche:

a) Beschreibung von Veränderungen im Zeitablauf
Es soll berechnet werden, wie groß die Wahrscheinlichkeit ist, bis zu einem gegebenen Zeitpunkt in einem bestimmten Zustand zu verbleiben (Überlebenswahrscheinlichkeit) bzw. in einem gegebenen Intervall den vorher eingenommenen Zustand zu wechseln (Sterbewahrscheinlichkeit). Darüber hinaus soll angegeben werden, mit welcher Rate Zustandswechsel (Ereignisse) auftreten. Diese drei charakteristischen Funktionen (oder monotone Transformationen derselben) sollen graphisch dargestellt werden, damit man ihren zeitlichen Verlauf besser beurteilen kann.

b) Gruppenvergleiche
Um erste Aussagen über mögliche Erklärungsfaktoren machen zu können, wird das Datenmaterial disaggregiert und untersucht, ob die unter a) berechneten Überlebensfunktionen und Raten sich zwischen Gruppen unterscheiden, die bezüglich bestimmter Merkmale homogen sind. Eine Möglichkeit, solche Gruppenunterschiede zu testen, ist die Berechnung von Konfidenzintervallen für die o.g. drei charakteristischen Funktionen. Damit sind jedoch nur zeitpunktspezifische Vergleiche möglich. Wünschenswert sind daher Testverfahren, die den ganzen Untersuchungszeitraum berücksichtigen.

c) Regressionsmodelle
Nachdem solchermaßen mögliche Erklärungsfaktoren und zeitliche Veränderungen des Prozesses exploriert wurden, will man konkrete Regressionsmodelle berechnen, in denen Raten auf einen Set von Kovariaten sowie unterschiedliche Formen zeitlicher Veränderung zurück-

geführt werden. Das bevorzugte Schätzverfahren ist Maximum– bzw. Partial Likelihood. Die Modelle unterscheiden sich eigentlich nur durch die Form der Zeitabhängigkeit. Die Palette reicht von unspezifizierten Veränderungen im Zeitablauf (COX's Regressionsmodell) über zeitkonstante Raten (Exponentialverteilung) bis hin zu komplizierten Funktionen der Zeit (u.a. Gompertz– oder Weibull–Verteilung). Die meisten Verfahren wurden zunächst für kontinuierliche Daten entwickelt, aber wie ALLISON (1982) zeigt, kann man bekannte statische Verfahren für diskrete Daten (z.B. logistische Regression) verwenden, um dynamische Modelle für diskrete Wartezeiten zu testen.

d) Modellevaluation
Hat man schließlich ein oder mehrere plausible Regressionsmodelle gefunden, dann möchte man wissen, wie gut sie die Daten beschreiben und wie realistisch die Modellannahmen sind. Es muß daher möglich sein,
   - bestimmte Modellannahmen zu überprüfen (z.B. die Annahme proportionaler Risiken) sowie
   - Modellprognosen und Daten miteinander zu vergleichen (Residuenanalyse).

   Erst auf dem Hintergrund dieser Tests läßt sich entscheiden, welches Modell unter den gegebenen Umständen das "beste" ist bzw. in welcher Weise die Modelle abzuändern wären.

Anhang D enthält eine Übersicht über die wichtigsten Programme zur Verlaufsdatenanalyse, die ganz grob nach diesen vier Aufgabenbereichen geordnet ist. Ausgewählte Programmbeispiele zu den Auswertungen in Kapitel 4 und 5 sind in Anhang E abgedruckt.

Die großen statistischen Programmpakete SAS, SPSS und BMDP haben dabei natürlich den Vorteil, daß sie weit verbreitet sind, ihre Anwendung daher vielen bekannt ist und sie außerdem nicht nur Verlaufsdatenanalyse können. Auf dem Hintergrund der Überlegungen im vorherigen Abschnitt und der Bedeutung des explorativen Teils der Verlaufsdatenanalyse kann man die Datenmanagementprobleme gar nicht hoch genug einschätzen. Hier erfährt man natürlich von den all–purpose Paketen sehr viel mehr Unterstützung als von stand–alone Programmen wie RATE oder TDA. Allerdings seien alle SPSS–Freunde, die an dieser Stelle vielleicht schon triumphieren, gewarnt, denn die Möglichkeiten der Verlaufsdatenanalyse beschränken sich bei SPSS im wesentlichen auf den explorativen Teil. Mit

der erheblichen Ausweitung des statistischen Leistungsumfangs ab Version 6.04 ist SAS der klare "Testsieger" bei den Paketen, vor allem auch wegen seiner exzellenten Datenmanagementfähigkeiten und seiner graphischen Ausgabemöglichkeiten.

Die beiden Programme RATE (TUMA 1980) und TDA (früher RATC, ROHWER 1991) sind speziell auf Verlaufsdatenanalyse zugeschnitten und haben dementsprechend ein sehr viel größeres Leistungsangebot statistischer Methoden. Im Gegensatz zu den Programmpaketen betrachten sie nicht Wartezeiten (duration analysis) sondern allgemeinere Prozesse mit verschiedenen Übergangsraten – daher auch die Namensgebung (RATE bzw. Transition Data Analysis). Als Standard–FORTRAN– bzw. C–Programme sind sie auf einer breiten Palette von Hardware–Plattformen ablauffähig. TDA läuft vor allem auch auf einem PC (vorzugsweise mit 80386–Prozessor). Die Übersicht in Anhang D kann natürlich nicht alle Statistik–Programme berücksichtigen, die auf dem Markt verfügbar sind.[1] Für Statistik–Spezialisten und Programmierer ist vielleicht noch erwähnenswert, daß Verlaufsdaten auch mit dem Programm GLIM (AITKEN et al. 1989: Kap. 6) und der Programmiersprache GAUSS (SCHNEIDER 1991) analysiert werden können. Weitere Hinweise auf verfügbare Software finden sich in allen einschlägigen Lehrbüchern (vgl. Einleitung).

1) LIMDEP und SYSTAT verfügen ebenfalls über Prozeduren zur Verlaufsdatenanalyse.

# 4. *Explorative Verfahren*

Jede Datenanalyse beginnt mit einfachen statistischen Verfahren, um einen Überblick über das Untersuchungsmaterial zu erhalten (explorative Datenanalyse), ehe konkrete Fragestellungen oder Hypothesen überprüft werden (konfirmatorische Datenanalyse). Bei Querschnittsdaten verwendet man hierzu einfache Häufigkeitsauszählungen, Kreuztabellen, Korrelations– und Assoziationsmaße sowie graphische Darstellungen (z.B. Histogramme, Streudiagramme etc.). Auch bei Verlaufsdaten geht es zunächst um eine Deskription des Datenmaterials. Dabei stehen folgende Fragen im Vordergrund:

- Wie hoch ist überhaupt das Ereignisrisiko in der Untersuchungsgruppe?
- Welche Untersuchungseinheiten sind in welcher Weise davon betroffen (heterogener Prozeß)?
- Wie verändert sich das Ereignisrisiko im Zeitablauf (zeitabhängiger Prozeß)?

Da unzureichend kontrollierte Heterogenität zu scheinbaren Zeitabhängigkeiten führen kann (vgl. Abschnitt 2.5.1), wird es notwendig sein, die beiden letzten Fragen im Zusammenhang zu erörtern.

Gesucht sind also Auswertungsverfahren, die eine Beschreibung der Stichprobe gestatten, ohne bestimmte Annahmen über den Verlaufsprozeß vorauszusetzen (*nicht–parametrische Verfahren*). Man verwendet dazu einfache deskriptive Techniken der Survival Analysis und durch Disaggregierung ist es möglich, die Heterogenität des Datenmaterials zu berücksichtigen (Methode des Gruppenvergleichs). Historisch gesehen stammen diese Verfahren aus der Demographie (Sterbetafeln), wurden dann auf biometrische (z.B. Krebsforschung) und technische Fragestellungen (z.B. Verschleißdauer von Maschinenteilen) übertragen und seit einiger Zeit findet man auch sozialwissenschaftliche Anwendungen. Bei den klassischen Anwendungen handelt es sich in der Regel um Prozesse mit singulären, nicht–wiederholbaren Ereignissen (z.B. Tod, Ausfall einer Maschine), so daß eine Übertragung auf sozialwissenschaftliche Fragestellungen mit multiplen und wiederholbaren Ereignissen mit gewissen Schwierigkeiten verbunden ist. Bei multiplen Ereignissen hilft man sich in der Regel dadurch, daß jeder Typ von Ereignis einzeln betrachtet wird und alle anderen Ereignisse als davon unabhängige Risiken zensiert werden. Da dieses eine Stan-

dardanwendung in den Sozialwissenschaften ist, beginnt dieses Kapitel mit einem Abschnitt 4.1 über konkurrierende Risiken und zensierte Beobachtungen.

In dieser Arbeit habe ich mich schon einmal mit der Analyse einfacher Wartezeitverteilungen beschäftigt (vgl. Abschnitt 2.2). Aus didaktischen Gründen habe ich dabei einen zeitdiskreten Prozeß unterstellt. In diesem Kapitel möchte ich alle vorliegenden Informationen ausschöpfen und die exakte Zeitdauer bis zum Eintritt eines Ereignisse berücksichtigen. Es werden zwei einfache Verfahren zur Schätzung der Überlebenswahrscheinlichkeit vorgestellt, wobei der Ansatz von KAPLAN und MEIER (Abschnitt 4.3) einige der Nachteile des Sterbetafelschätzers (Abschnitt 4.2) umgeht. Der Sterbetafelansatz verwendet gruppierte Wartezeiten, d.h. die vorliegenden Zeitdauern werden nachträglich klassifiziert. Der Ansatz von KAPLAN und MEIER verwendet dagegen direkt die Wartezeiten.

In Abschnitt 4.4 beschäftige ich mich dann mit der Frage, wie man durch die Methode des Gruppenvergleichs die Heterogenität des Datenmaterials kontrolliert und erste Aufschlüsse über mögliche Erklärungsfaktoren erhält. An dieser Stelle wird die Unterscheidung zwischen explorativen und konfirmatorischen Verfahren brüchig, denn die entsprechenden Tests lassen sich natürlich auch wie klassische Hypothesentests verwenden. Im Rahmen eines Experiments könnte man solchermaßen die Unterschiede zwischen Experimental– und Kontrollgruppe überprüfen.

Nachdem die Stichprobe in (möglichst) homogene Subpopulationen unterteilt ist, kann man scheinbare Zeitabhängigkeiten weitgehend ausschließen. Auf dieser Stufe der Datenanalyse bietet es sich dann an, die Übereinstimmung der Daten mit verschiedenen Verteilungsmodellen für Wartezeiten zu untersuchen. In Abschnitt 4.5 zeige ich, wie man mit Hilfe graphischer Darstellungen überprüft, ob die Daten eher durch einen Prozeß mit konstanter oder eher durch einen Prozeß mit zeitabhängiger Rate beschrieben werden.

Dieses Kapitel wird durch einen Abschnitt 4.6 abgeschlossen, in dem ich noch einmal die Analyse multipler und wiederholbarer Ereignisse aufgreife und verschiedene Erweiterungen des o.g. Standardvorgehens beschreibe.

## 4.1 *Konkurrierende Risiken und zensierte Beobachtungen*

In Tabelle 2.1 ist die Entwicklung der Risikomenge, die für die Berechnung der Überlebenswahrscheinlichkeit und der Rate grundlegend ist, nur eine Funktion der stattfindenden Ereignisse. In allen real vorkommenden Verlaufsanalysen wird man jedoch noch Abgänge anderer Art zu verzeichnen haben. Auf einige typische Fälle dieser weiteren Abgänge möchte ich an Hand des Mobilitätsbeispiels kurz eingehen. Die folgende Abbildung 4.1 zeigt dazu für eine Reihe von Personen, wie lange die jeweils erste Tätigkeit ausgeübt wurde (Tätigkeitsdauer) und wie lange die jeweilige Person an der Untersuchung teilnahm (Beobachtungsdauer)[1]. Diese Verlaufsdaten können *prozeßbegleitend oder retrospektiv erhoben* worden sein. Abbildung 4.1 zeigt also praktisch einen Ausschnitt aus Abbildung 1.4.

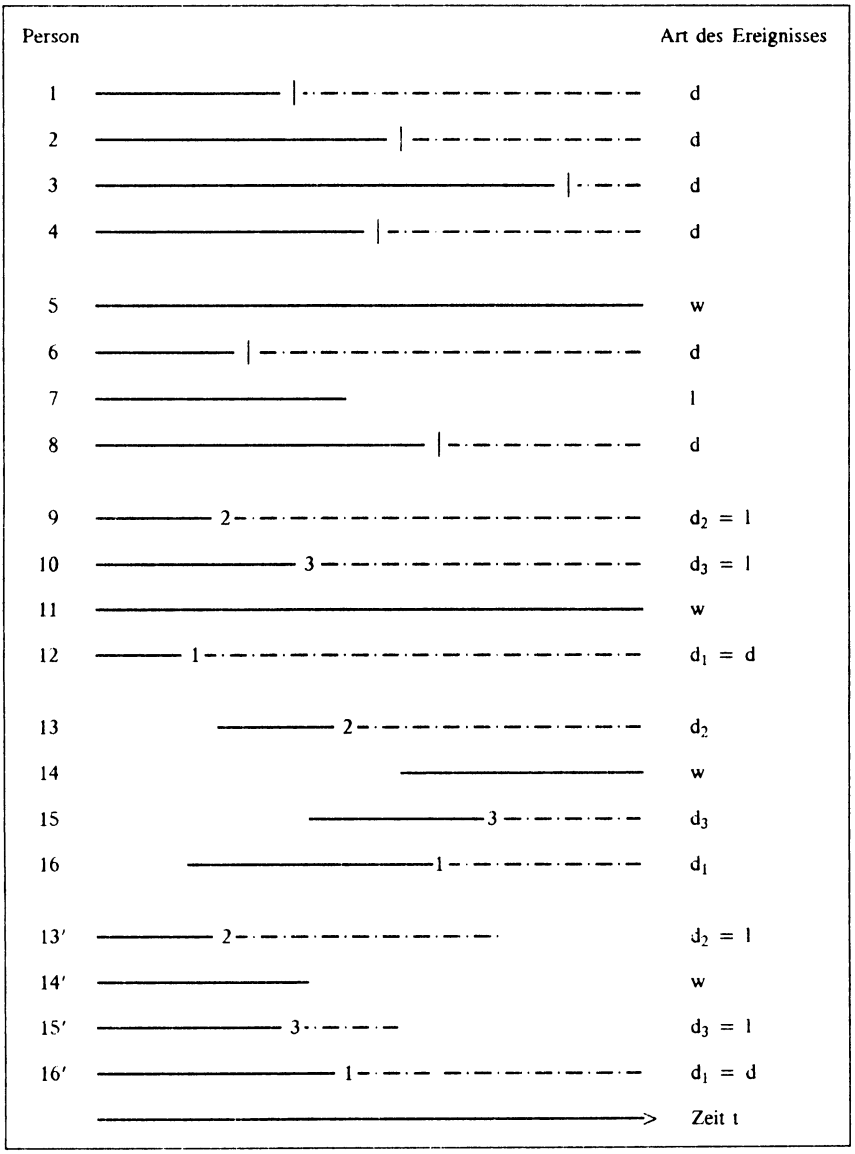
Die ersten vier Personen könnten aus der Stichprobe in Tabelle 2.1 stammen. Alle wurden über einen gleich langen Zeitraum beobachtet, der außerdem so lang war, daß jede Person mindestens ein Ereignis zu verzeichnen hatte. Der ursprüngliche Stichprobenumfang (Risikomenge) verringert sich also sukzessive, weil Personen wegen eines Ereignisses ausscheiden. Alle Verläufe werden daher mit einem d (=dead) gekennzeichnet.

Die folgenden Personen 5−8 charakterisieren eine andere Untersuchung, die im Prinzip ähnlich lange dauert wie die vorhergehende, in der jedoch zusätzliche Komplikationen auftreten. Hier wird eine Kohorte von Personen prospektiv betrachtet, aber nicht jede Person hat im Untersuchungszeitraum ein Ereignis (vgl. Person 5) und außerdem sind einzelne im Zeitablauf nicht mehr auffindbar (vgl. Person 7). Person 5 beendet also die Untersuchung ohne Ereignis (w = withdrawn alive), während der Kontakt zu Person 7 im Verlauf der Untersuchung abreißt (l = lost). Da hier eine Kohorte von Personen betrachtet wird, kann das "Ereignis" w nur am Ende der Untersuchungsperiode, das Ereignis l hingegen jederzeit auftreten.

Die Personen 9−12 könnten aus einer Kohorte von Berufsanfängern stammen, die alle zum gleichen Zeitpunkt in das Erwerbsleben eingetreten

1) Um auf die einzelnen Fälle besser Bezug nehmen zu können, sind die Personen durchnumeriert. Dies soll jedoch nicht heißen, daß sie alle aus der gleichen Untersuchung stammen.

*Abbildung 4.1: Typische Abgänge im Rahmen einer Verlaufsanalyse*




Zeichenerklärung: —— Tätigkeitsdauer —·· Beobachtungszeitraum
| Tätigkeitswechsel, 1 Abstieg, 2 Aufstieg, 3 horizontale Mobilität

sind und über ihre ersten 10 Berufsjahre befragt werden. Wenn die Erhebung retrospektiv erfolgt, können keine Ausfälle von Untersuchungspersonen auftreten. Einige Personen (z.B. Person 11) können aber die Untersuchung ohne Ereignis beenden, weil innerhalb der ersten 10 Berufsjahre kein Tätigkeitswechsel auftrat. Zusätzlich sollen jetzt verschiedene Arten von Ereignissen auftreten (Ab− und Aufstiege, horiz. Mobilität: $d_1$, $d_2$, $d_3$). Angenommen man interessiert sich für das Auftreten beruflicher Abstiege, dann hat Person 12 ein Ereignis d, während Person 11 die Untersuchung ohne "Ereignis" w beendet und die Personen 9 und 10 aus anderen Gründen l ausscheiden.

Bei den bisher diskutierten Fällen kann das "Ereignis" w nur bei den jeweils längsten Wartezeiten auftreten, da es sich um Kohortendaten handelt. Nicht so bei den Personen 13−16: Sie charakterisieren eine Erhebung mit variierenden Eintrittsdaten, wie sie z.B. auftreten, wenn man die vorher diskutierten Berufsanfänger zwar über einen gleich langen Zeitraum untersucht, diese Personen aber zu unterschiedlichen Zeitpunkten in das Berufsleben eintreten. In dieser Form können die Daten nicht ausgewertet werden, sie müssen erst an Hand ihres Eintrittsdatums synchronisiert werden (vgl. Personen 13'−16'). Die Zeit wird jetzt nicht mehr chronologisch, sondern in Jahren seit Berufseintritt gemessen. Mit dieser neuen Zeitdefinition kann das "Ereignis" w auch schon bei kürzeren Wartezeiten auftreten (vgl. Person 14'). Darüber hinaus gibt es wie bei der vorhergehenden Untersuchung verschiedene Arten von Ereignissen, die man je nach Forschungsinteresse in primäre d und konkurrierende Risiken l unterteilen kann.

Im Gegensatz zu den Daten in Tabelle 2.1 (vgl. Personen 1−4) verringert sich die Risikomenge bei den drei folgenden Untersuchungen aus verschiedenen Gründen: Es treten Ereignisse d auf, darüber hinaus gibt es aber noch Abgänge ohne Ereignis w und aus anderen Gründen l. Das "Ereignis" w entspricht den bereits in den vorhergehenden Kapiteln besprochenen zensierten Beobachtungen auf Grund einer begrenzten Untersuchungsdauer. Das Ereignis l entspricht den in Abschnitt 2.4.1 besprochenen konkurrierenden Risiken. Beide Ausgangsdaten w und l werden nun üblicherweise als zensierte Beobachtungen $c = w + l$ zusammengefaßt. Dabei drängt sich natürlich die Frage auf, ob es sinnvoll ist, die Ereignisse w und l gleichermaßen als zensierte Beobachtungen zu betrachten. Offensichtlich kommen beide Ereignisse doch aus ganz verschiedenen Gründen zustande. Das "Ereignis" w ist häufig eine externe, vom Forscher geplante Beschränkung der Untersuchungsdauer, die einzelne Verläufe einfach abschneidet

(*geplante Zensierung*). Das Ereignis 1 hingegen ist ein immanentes Ergebnis des untersuchten Prozesses, das eigentlich erklärt werden sollte (*ungeplante Zensierung*). Hintergrund dieses Vorgehens ist jedoch die Behandlung zensierter Beobachtungen als eines unter mehreren konkurrierenden Risiken (vgl. Abschnitt 2.5.2.2). Für eine einfache statistische Weiterverarbeitung muß lediglich vorausgesetzt werden, daß die verschiedenen konkurrierenden Risiken – also auch die w zensierten Beobachtungen im engeren Sinne – von dem eigentlich interessierenden (primären) Risiko unabhängig sind. Diese Annahme ist bei geplanten Zensierungen der beschriebenen Art plausibel. Für die ungeplanten Zensierungen ist diese Annahme im Einzelfall zu prüfen. Wenn wir im folgenden beide Gruppen zusammenfassen, dann gehen wir implizit von der Gültigkeit dieser Annahme aus.

## 4.2 *Nicht–parametrische Verfahren für gruppierte Wartezeiten*

Zunächst einmal stellt sich die Frage, warum man eigentlich gruppierte Wartezeiten verwendet, wenn man doch alle vorliegenden Informationen, also die exakten Wartezeiten ausschöpfen möchte. Diese Frage läßt sich auf verschiedene Weise beantworten: Erstens ist es übliche Praxis, kontinuierliche Merkmale zunächst einmal zu klassifizieren, um einen Überblick über ihre Verteilungsform zu erhalten. Zweitens liegen viele Verlaufsdaten schon in aggregierter Form vor, weil sie entweder zeitdiskret oder ungenau erhoben wurden (vgl. die MZ71–Daten). In diesem Fall wäre mit der Klassifikation gar kein Informationsverlust verbunden. Drittens kann es bei sehr umfangreichen Datensätzen EDV–technisch effizienter sein, nicht die Originaldaten zu verwenden. Schließlich gibt es in der Demographie schon seit langer Zeit das Instrument der Sterbetafelanalyse, das für die Auswertung von Verlaufsdaten wie geschaffen ist.

Sterbetafeln werden in der Demographie dazu verwendet, um mit Hilfe von Querschnittsdaten Aussagen über die Mortalität verschiedener Altersgruppen zu machen. Dabei werden die Sterbefälle einer Altersklasse auf den mittleren (jährlichen) Bevölkerungsbestand derselben Altersklasse bezogen. Da es sich um Querschnittsdaten handelt, existiert das Problem zensierter Beobachtungen nicht. Will man daher Verlaufsdaten mit Sterbetafeln analysieren, muß dieser Ansatz entsprechend verallgemeinert werden. Man spricht daher auch von verallgemeinerten Sterbetafeln.

### 4.2.1 *Vorgehen bei Sterbetafelschätzungen*

Ich möchte das Vorgehen an Hand der MZ71−Daten illustrieren und verwende dazu die Daten aus Tabelle 3.2 (PRG0401).[1] Es soll um die Frage gehen, wie sich die sozialen Abstiege in den ersten 10 Berufsjahren verteilen. Ich werde also das Instrument der Sterbetafelanalyse benutzen, um die charakteristischen Funktionen zur Beschreibung der Wartezeitverteilung bis zu einem Abstieg zu schätzen. Im Gegensatz zum Vorgehen in Abschnitt 2.2 soll nicht davon ausgegangen werden, daß sich alle Ereignisse auf einen Zeitpunkt (das Intervallende) konzentrieren. Stattdessen wird angenommen, daß sich Ereignisse über das gesamte Intervall verteilen, was für das MZ71−Beispiel auch die realistischere Annahme sein dürfte.

Eine *verallgemeinerte Sterbetafel*, enthält im einzelnen folgende Einträge:

a) Intervallbeginn $\tau_{k-1}$ und Intervallende $\tau_k$: Alle Wartezeiten werden in $k=1,2,\ldots,K+1$ Intervalle $[\tau_{k-1},\tau_k)$ eingeordnet, die jeweils zum Zeitpunkt $\tau_{k-1}$ beginnen ($\tau_0=0$). Diese Intervalle werden als gegeben angenommen, so daß die Zahl der Ereignisse, die in das jeweilige Intervall fällt, die interessierende (multinomial verteilte) Zufallsvariable ist.

b) Intervallmittelpunkte $\tau_{mk}$: Diese Klassenmittelpunkte dienen der späteren Zeichnung der verschiedenen Funktionen.

c) Intervallbreite $\Delta\tau_k=\tau_k-\tau_{k-1}$: Je mehr Zeitpunkte das jeweilige Intervall umfaßt, um so mehr Ereignisse sind innerhalb dieses Zeitraums wahrscheinlich. Da jetzt die Intervalldauer berücksichtigt werden soll, wird $\Delta\tau_k$ zur Berechnung der einzelnen Funktionen benötigt. Jedes Intervall dauert vom Zeitpunkt $\tau_{k-1}$ bis (aber nicht einschließlich) zum Zeitpunkt $\tau_k$. Insgesamt umfaßt es $\Delta\tau_k$ Zeiteinheiten, wobei das letzte Intervall theoretisch nach hinten unbegrenzt ist ($\tau_{K+1}=\infty$), so daß die Intervalldauer unendlich und eine entsprechende Berechnung einzelner Funktionen nicht sinnvoll ist.

---

1) Die entsprechenden Computer−Programme sind in Anhang E dokumentiert. Der Name dieses Programms ist PRG0401. Es kann an Hand dieses Kürzels im Anhang gefunden werden. Alle weiteren Programme werden im Text in der gleichen Form (PRG−Kapitel−Nummer) zitiert.

d) Zahl der in das Intervall eintretenden Untersuchungseinheiten $n_k^*$: Die Zahl der in das erste Intervall eintretenden Untersuchungseinheiten $n_1^*$ entspricht dem Stichprobenumfang N. Alle folgenden Zahlen $n_k^*$ ergeben sich aus dem Wert der Vorperiode $n_{k-1}^*$ abzüglich der Abgänge l, w und d:

$$n_k^* = n_{k-1}^* - l_{k-1} - w_{k-1} - d_{k-1} \tag{4.1}$$

e) Zahl der Untersuchungseinheiten ohne Ereignis $w_k$.

f) Zahl der Untersuchungseinheiten $l_k$, die aus anderen Gründen ausscheiden.

g) Zahl der Untersuchungseinheiten mit Ereignis $d_k$.

Da es jetzt um eine Analyse beruflicher Abstiege geht, sind in der Spalte $l_k$ alle anderen Tätigkeitswechsel (Aufstiege, horiz. Mobilität) zusammengefaßt. Zusammen mit der Gruppe $w_k$ ergeben sie die Anzahl $c_k$ der pro Intervall zensierten Beobachtungen. Die spannende Frage ist nun: Wie groß ist die Risikomenge pro Intervall, wenn ein Teil $c_k$ der in das Intervall eintretenden Personen $n_k^*$ vor Intervallende ausscheidet? Die Sterbetafelanalyse nimmt hier üblicherweise an, daß sich die Ereignisse $l_k$ und $w_k$ gleichmäßig über die Intervalldauer verteilen (vgl. BERKSON/GAGE 1950, CUTLER/ EDERER 1958). Mit anderen Worten, die Personen $l_k$ und $w_k$ hatten durchschnittlich die Hälfte der Intervalldauer das Risiko, einen beruflichen Abstieg zu erfahren. Hinter diesem Argument stehen keine großen statistischen Ableitungen. Es ist eine mehr oder weniger plausible ad-hoc-Begründung für die folgende Berechnung der Risikomenge:

h) Risikomenge $n_k$: Auf Grund der Annahme einer während des Intervalls gleichverteilten Zensierung ergibt sich die Risikomenge aus der Anzahl der in das Intervall eintretenden Untersuchungseinheiten $n_k^*$ abzüglich der Hälfte der zensierten Beobachtungen $c_k$.

$$n_k = n_k^* - \frac{c_k}{2} = n_k^* - \frac{l_k + w_k}{2} \tag{4.2}$$

Alle diese Daten sind in der folgenden Sterbetafel 4.1 festgehalten. Sie sind der Ausgangspunkt für die folgenden Schätzungen der Überlebenswahrscheinlichkeit. Die Berechnungen beruhen dabei auf den in Abschnitt 2.2 besprochenen Formeln.

*Tabelle 4.1: Eine verallgemeinerte Sterbetafel zur Analyse beruflicher Abstiege*

| k | $\tau_{k-1}$ | $\tau_k$ | $\Delta\tau_k$ | $n_k^*$ | $w_k$ | $l_k$ | $d_k$ | $n_k$ | $\hat{q}_k$ | $\hat{p}_k$ | $\hat{S}(\tau_{mk})$ | $\hat{f}(\tau_{mk})$ | $\hat{f}(\tau_{mk})$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 10666 | 0 | 0 | 0 | 10666,0 | 0,0000 | 1,0000 | 1,0000<br>(0,000) | 0,0000<br>(0,000) | 0,0000<br>(0,000) |
| 2 | 1 | 2 | 1 | 10666 | 0 | 299 | 129 | 10516,5 | 0,0123 | 0,9877 | 1,0000<br>(0,001) | 0,0123<br>(0,000) | 0,0123<br>(0,001) |
| 3 | 2 | 3 | 1 | 10238 | 0 | 398 | 126 | 10039,0 | 0,0126 | 0,9874 | 0,9877<br>(0,002) | 0,0126<br>(0,001) | 0,0124<br>(0,001) |
| 4 | 3 | 4 | 1 | 9714 | 0 | 444 | 170 | 9492,0 | 0,0179 | 0,9821 | 0,9753<br>(0,002) | 0,0181<br>(0,001) | 0,0175<br>(0,001) |
| 5 | 4 | 5 | 1 | 9100 | 0 | 434 | 114 | 8883,0 | 0,0128 | 0,9872 | 0,9579<br>(0,002) | 0,0129<br>(0,001) | 0,0123<br>(0,001) |
| 6 | 5 | 6 | 1 | 8552 | 0 | 429 | 119 | 8337,5 | 0,0143 | 0,9857 | 0,9456<br>(0,003) | 0,0144<br>(0,001) | 0,0135<br>(0,001) |
| 7 | 6 | 7 | 1 | 8004 | 0 | 446 | 99 | 7781,0 | 0,0127 | 0,9873 | 0,9321<br>(0,003) | 0,0128<br>(0,001) | 0,0119<br>(0,001) |
| 8 | 7 | 8 | 1 | 7459 | 0 | 482 | 125 | 7218,0 | 0,0173 | 0,9827 | 0,9202<br>(0,003) | 0,0175<br>(0,001) | 0,0159<br>(0,001) |
| 9 | 8 | 9 | 1 | 6852 | 0 | 528 | 143 | 6588,0 | 0,0217 | 0,9783 | 0,9043<br>(0,003) | 0,0219<br>(0,002) | 0,0196<br>(0,002) |
| 10 | 9 | 10 | 1 | 6181 | 1638 | 369 | 115 | 5177,5 | 0,0222 | 0,9778 | 0,8847<br>(0,004) | 0,0225<br>(0,002) | 0,0196<br>(0,002) |
| 11 | 10 | 11 | 1 | 4059 | 1600 | 344 | 65 | 3087,0 | 0,0211 | 0,9789 | 0,8650<br>(0,004) | 0,0213<br>(0,002) | 0,0182<br>(0,002) |
| 12 | 11 | $\infty$ | $\infty$ | 2050 | 2050 | 0 | 0 | 1025,0 | 0,0000 | 1,0000 | 0,8469<br>(0,003) | – | – |

Angaben in Klammern: Standardfehler

Wegen der Korrektur (4.2) der Risikomenge um die zensierten Beobachtungen in jedem Intervall ist eine direkte Anwendung von Gleichung (2.6) nicht möglich. Es verbleibt lediglich die Möglichkeit, die Überlebenswahrscheinlichkeit durch sukzessive Multiplikation zu errechnen (vgl. Gleichung 2.15). Dafür definiert man für jedes Intervall k eine bedingte Sterbewahrscheinlichkeit, die angibt, mit welcher Wahrscheinlichkeit ein Ereignis im Intervall k auftritt, vorausgesetzt alle vorhergehenden Intervalle wurden überlebt. Ein Schätzer für diese bedingte Sterbewahrscheinlichkeit $q_k$ ist:

$$(4.3) \qquad \hat{q}_k = \begin{cases} \dfrac{d_k}{n_k} = \dfrac{d_k}{n_k^* - \dfrac{c_k}{2}} & \text{für } k = 1, \ldots, K \\ 1 & \text{für } k = K+1 \end{cases}$$

Dementsprechend ergibt sich eine bedingte Überlebenswahrscheinlichkeit $p_k = 1 - q_k$ und durch sukzessive Multiplikation ein Schätzer für die (unbedingte) *Überlebenswahrscheinlichkeit* zu Beginn $\tau_{k-1}$ des k−ten Intervalls[1]:

$$(4.4) \qquad \hat{S}(\tau_{k-1}) = \prod_{i=0}^{k-1} \hat{p}_i = \prod_{i=0}^{k-1} (1-\hat{q}_i) \quad \text{mit } \hat{p}_0 = 1,\ \hat{q}_0 = 0$$

Gleichung (4.4) bezeichnet man als den sogenannten *Sterbetafelschätzer* der Überlebenswahrscheinlichkeit. In einigen Fällen ist es auch interessant, Dichte f(t) und Rate r(t) für gruppierte Wartezeiten empirisch zu bestimmen. Zwei entsprechende Schätzer, die von (4.4) abgeleitet werden können, lauten:

$$(4.5) \qquad \hat{f}(\tau_{mk}) = \frac{\hat{S}(\tau_k) - \hat{S}(\tau_{k+1})}{\Delta\tau_k} = \frac{\hat{S}(\tau_k)\,\hat{q}_k}{\Delta\tau_k}$$

$$(4.6) \qquad \hat{r}(\tau_{mk}) = \frac{\hat{f}(\tau_{mk})}{\hat{S}(\tau_{mk})} = \frac{\hat{f}(\tau_{mk})}{\dfrac{\hat{S}(\tau_{mk}) + \hat{S}(\tau_{mk+1})}{2}} = \frac{2\,\hat{q}_k}{\Delta\tau_k\,(2-\hat{q}_k)} = \frac{2\,\hat{q}_k}{\Delta\tau_k\,(\hat{p}_k+1)}$$

(4.5) geht davon aus, daß f(t) innerhalb eines Intervalls der Dichte einer Gleichverteilung entspricht. In diesem Fall ist die Wahrscheinlichkeit $P(\tau_{k-1} \le T < \tau_k)$ eines Ereignisses innerhalb des k−ten Intervalls $\Delta\tau_k \cdot f(t)$. Schätzt man die Wahrscheinlichkeit durch $\hat{S}(\tau_k) - \hat{S}(\tau_{k+1})$, dann ergibt sich schließlich (4.5). Gleichung (4.6) kombiniert analog (2.11) Dichte und Überlebenswahrscheinlichkeit, um die Rate zu schätzen. Da r(t) für die

1) Die SURVIVAL−Prozedur des Programmpaketes SPSS berechnet die Überlebenswahrscheinlichkeit $S(\tau_k)$ am Ende $\tau_k$ des k−ten Intervalls. Dementsprechend ist über alle Intervalle bis einschließlich k zu multiplizieren (i = 0,1,...k).

Intervallmitte bestimmt werden soll, wird die Überlebenswahscheinlichkeit gemäß der Gleichverteilungannahme interpoliert.

Tabelle 4.1 zeigt die Schätzergebnisse für S(t), r(t) und f(t) inkl. Standardfehler. Am Wert der Überlebenswahrscheinlichkeit zu Beginn des 11. Intervalls erkennt man, daß 86,5% der Stichprobe die ersten 10 Berufsjahre ohne Abstieg überstanden hat. Wenn überhaupt, dann sind berufliche Abstiege im 9. und 10. Intervall am wahrscheinlichsten, denn in diesen Intervallen, d.h. nach 8−10 Berufsjahren, verringert sich die Überlebenswahrscheinlichkeit am stärksten (vgl. auch den Dichteschätzer in der letzten Spalte).

Eine weitere nützliche Statistik zur Beschreibung der Sterbetafel ist der Median der Wartezeiten, d.h. die Zeitdauer, nach der genau die Hälfte der Stichprobe mindestens ein Ereignis hatte (*mittlere Zustandsdauer*)[1]. Man bestimmt zunächst an Hand der Überlebenswahrscheinlichkeit das Intervall j, an dessen Endpunkt mindestens 50% der Stichprobe ein Ereignis hatte: $S(\tau_j) > 0{,}5$ und $S(\tau_{j+1}) \leq 0{,}5$. Der Median $\tilde{t}$ liegt dann zwischen Intervallbeginn und −ende und läßt sich näherungsweise durch lineare Interpolation bestimmen:

$$(4.7) \qquad \tilde{t} = \tau_j + \Delta\tau_j \frac{0{,}5 - \hat{S}(\tau_j)}{\hat{S}(\tau_j) - \hat{S}(\tau_{j+1})} = \tau_j + \Delta\tau_j \frac{0{,}5 - \hat{S}(\tau_j)}{\hat{f}(\tau_{mj})}$$

Statt des 50. (d.h. dem Median) kann man natürlich auch jedes andere Perzentil verwenden. In dem Anwendungsbeispiel ist es z.B. sinnvoll, das 10. Perzentil zu berechnen, da der Prozeß nicht ausreichend lange untersucht wurde, um bei mindestens 50% der Stichprobe einen Abstieg zu beobachten. Danach hat erst nach 8,78 Jahren mindestens 10% der Stichprobe einen beruflichen Abstieg erfahren (zum Vergleich berufliche Aufstiege: 4,62 Jahre). Bevor ich mit der allgemeinen Diskussion fortfahre, möchte ich mich mit den Eigenschaften des Sterbetafelschätzers und alternativen Schätzverfahren etwas genauer beschäftigen.

---

1) Die Berechnung eines arithmetischen Mittelwertes ist durch die Tatsache erschwert, daß die gesamte Verteilung nur unvollständig beobachtet werden kann. Dagegen liegen für einen Teil der Verteilung (Perzentil) immer genügend vollständige Beobachtungen vor. Wenn man aber dennoch einen Mittelwert berechnen möchte, dann sollte man von exakten statt von gruppierten Wartezeiten ausgehen (vgl. Abschnitt 4.3.1).

### 4.2.2 *Annahmen der Sterbetafelschätzungen*

Die Ableitung des Sterbetafelschätzers erfolgte mehr oder weniger nach Plausibilitätsüberlegungen. In diesem Abschnitt möchte ich das Problem unter mehr formalstatistischen Gesichtspunkten betrachten und mich mit den impliziten Annahmen dieses Vorgehens beschäftigen. Es sind verschiedene alternative Schätzverfahren für gruppierte Wartezeiten vorgeschlagen worden, und ich werde kurz auf diese Debatte eingehen.

Angenommen $n_k^*$ Personen beginnen das Intervall k. Davon überstehen $n_{k+1}^*$ Personen die gesamte Intervalldauer $\Delta\tau_k$, während der Rest $(w_k + l_k + d_k)$ ausscheidet. Die Ausgangsinformation besteht also aus $d_k$ unzensierten und $(n_{k+1}^* + l_k + w_k)$ zensierten Wartezeiten, deren Gesamtdauer $U_k$ bzw. $V_k$ Zeiteinheiten beträgt. Wenn man nun annimmt, daß Ereignisse im Intervall mit konstanter Rate $\lambda_k$ auftreten, dann ist analog (2.60) der Quotient aus Ereignissen $d_k$ und Gesamtdauer $(U_k + V_k)$ aller Wartezeiten ein unverzerrter Schätzer der intervallspezifischen Rate:

$$\hat{\lambda}_k = \frac{d_k}{U_k + V_k} \tag{4.8}$$

Die Gesamtdauern $U_k$ und $V_k$ können durch Summen individueller Wartezeiten ausgedrückt werden:

$$U_k = \sum_i^{d_k} (t_i^d - \tau_{k-1})$$

$$V_k = \sum_i^{w_k} (t_i^w - \tau_{k-1}) + \sum_i^{l_k} (t_i^l - \tau_{k-1}) + n_{k+1}^* \Delta\tau_k$$

$\tau_{k-1}$ steht für Intervallbeginn und $t_i$ bezeichnet die Wartezeit des Individuums i, die im Intervall k endet, wobei die Art der Beendigung durch einen hochgestellten Buchstaben d, w oder l gekennzeichnet wird. Bei gruppierten Wartezeiten sind jedoch die exakten Zeitpunkte $t_i$ nicht mehr zu erkennen. Wenn man nun wie in der Sterbetafelanalyse annimmt, daß die Zeitstrecke $(t_i - \tau_{k-1})$ im Mittel der halben Intervallbreite entspricht, läßt sich (4.8) weiter vereinfachen:

(4.9) $$\hat{\lambda}_k = \frac{d_k}{\Delta\tau_k \left( \frac{d_k}{2} + \frac{w_k}{2} + \frac{l_k}{2} + n^*_{k+1} \right)} ,$$

wegen $U_k = d_k \frac{\Delta\tau_k}{2}$ und $V_k = w_k \frac{\Delta\tau_k}{2} + l_k \frac{\Delta\tau_k}{2} + n^*_{k+1} \Delta\tau_k$ .

Mit $n^*_{k+1} = n_k - l_k/2 - w_k/2 - d_k$ ergibt sich schließlich Formel (4.6):

$$\hat{\lambda}_k = \frac{d_k}{\Delta\tau_k \left( n_k - \frac{d_k}{2} \right)} = \frac{2\frac{d_k}{n_k}}{\Delta\tau_k \left( 2 - \frac{d_k}{n_k} \right)} = \frac{2\,\hat{q}_k}{\Delta\tau_k (2 - \hat{q}_k)}$$

Wie sinnvoll ist jedoch die ad-hoc-Annahme $(t_i - \tau_{k+1}) = \Delta\tau_k/2$?

Angenommen Ereignisse $d_k$ und zensierte Beobachtungen $(l_k + w_k)$ treten im Intervall mit der gleichen konstanten Rate auf und können daher wie ein und dasselbe Ereignis behandelt werden. Unter welchen Bedingungen ist der Durchschnitt dieser Zeitstrecken gleich der Hälfte der Intervalldauer? Gesucht ist also der Mittelwert aller Wartezeiten mit Ereignis in einem Prozeß begrenzter Dauer $\Delta\tau = \Delta\tau_k$. Nach (2.58) ergibt sich der Erwartungswert eines solchen Poisson-Prozesses wie folgt aus der angenommenen zeitkonstanten Rate $\lambda$:

$$E(t|\text{Ereignis}, t < \Delta\tau) = \lambda^{-1} \frac{1 - (\lambda\Delta\tau + 1)\exp(-\lambda\Delta\tau)}{1 - \exp(-\lambda\Delta\tau)}$$

Mit dieser Gleichung kann man für Intervalle unterschiedlicher Länge und für verschiedene Raten die entsprechenden Erwartungswerte ausrechnen.

Wie man sieht, ist die Hälfte der Intervalldauer für verschiedene $\lambda$-Werte dann ein relativ guter Schätzer für den Erwartungswert, wenn man kurze Intervalle betrachtet. Je größer jedoch das Intervall wird, um so weniger paßt diese ad-hoc-Annahme. Es ist daher zu erwarten, daß die Eigenschaften des Sterbetafelschätzers um so schlechter werden, je länger die Intervalldauer ist.

*Tabelle 4.2: Mittelwert exponentiell verteilter Wartezeiten mit Ereignis in einem Prozeß begrenzter Dauer*

| $\lambda$ | $\Delta\tau = 2$ | $\Delta\tau = 5$ | $\Delta\tau = 10$ |
|---|---|---|---|
| 0,000001 | 1,000 | 2,600 | 5,100 |
| 0,00001 | 0,510 | 2,502 | 5,000 |
| 0,0001 | 0,500 | 2,500 | 4,999 |
| 0,001 | 0,500 | 2,499 | 4,992 |
| 0,005 | 0,500 | 2,490 | 4,958 |
| 0,01 | 0,499 | 2,479 | 4,917 |
| 0,05 | 0,496 | 2,396 | 4,585 |
| 0,1 | 0,492 | 2,293 | 4,180 |
| 0,5 | 0,459 | 1,553 | 1,932 |
| 1,0 | 0,418 | 0,966 | 1,000 |
| 5,0 | 0,193 | 0,200 | 0,200 |

Für dieses Rechenbeispiel mußte ich zwei Annahmen machen: a) exponentiell verteilte Wartezeiten mit intervallspezifischer Rate, b) gleiche Verteilung von Ereignissen und zensierten Beobachtungen. Andere Annahmen sind zumindest ebenso plausibel. Je nachdem, welches Verteilungsmodell man für Ereignisse und zensierte Beobachtungen unterstellt, kann man alternative Schätzer für die bedingte Sterbewahrscheinlichkeit $q_k$ entwickeln. Ein Überblick über die verschiedenen Ansätze findet sich bei ELANDT-JOHNSON/JOHNSON (1980: 162ff., insbes. Tabelle 6.3). Es zeigt sich, daß der *Sterbetafelschätzer* die tatsächliche Überlebenswahrscheinlichkeit unterschätzt: Die Verzerrung ist um so größer, je niedriger die Überlebenswahrscheinlichkeit und je länger die Intervalldauer ist (LITTELL 1952). Alle anderen Schätzer ergeben jedoch nahezu identische Resultate, wenn $n_k$ groß und $q_k$ klein ist. Da der Sterbetafelschätzer von allen der einfachste ist, empfehlen ELANDT-JOHNSON/JOHNSON seine Verwendung: "It is a good and robust estimator of $q_k$" (1980: 172). In jedem Fall ist die Schätzung aber davon abhängig, wie die Intervallgrenzen gewählt werden. Darüber entscheidet der Forscher nach mehr oder weniger willkürlichen Kriterien.

Abschließend noch ein paar Bemerkungen zur Zensierung von Aufstiegen und horizontaler Mobilität. Es kann gezeigt werden, daß bei Verwendung exakter Wartezeiten die *Risiken multipler Ereignisse unabhängig voneinander geschätzt* werden können, wenn die jeweils nicht interessierenden Ereignisse von den zu untersuchenden Ereignissen unabhängig sind und wie zensierte Beobachtungen behandelt werden (KALBFLEISCH/PRENTICE

1980: 168ff., s.a. Abschnitt 5.2.6). Natürlich hätte man auch soziale Aufstiege als primär interessierendes Ereignis wählen können. Dann hätte man die anderen Arten von Tätigkeitswechseln zensiert. Man kann zeigen, daß die solchermaßen geschätzten Einzelrisiken in der Summe das Gesamtrisiko ergeben, überhaupt irgendeinen Tätigkeitswechsel zu erfahren (vgl. Gleichung 2.30).

Möchte man also multiple Ereignisse untersuchen, dann bleibt einem nichts anderes übrig, als die Analyse in mehreren Schritten durchzuführen, in denen jeweils eines der multiplen Ereignisse betrachtet wird und die verbleibenden wie zensierte Beobachtungen behandelt werden. Das ist zugegebenermaßen etwas umständlich.

Bei dem folgenden Kaplan–Meier–Schätzer verfährt man genauso. Auch hier kann man zeigen, daß das Produkt der Überlebenswahrscheinlichkeiten für jedes Einzelereignis der Gesamtwahrscheinlichkeit entspricht, alle verschiedenen Ereignisse zu überleben, vorausgesetzt es treten keine Ereignisse zum gleichen Zeitpunkt auf (ties) (vgl. LAWLESS 1982: 486ff.).

## *4.3 Nicht–parametrische Verfahren für exakte Wartezeiten*

Wenn man exakte statt gruppierte Wartezeiten verwendet, ist eine Annahme über die Verteilung der Ereignisse pro Intervall nicht notwendig. Bei Verwendung der genauen Zeitdauern bis zum Eintritt der Ereignisse d, l und w in Gleichung (4.8) ergibt sich für exponentiell verteilte Wartezeiten ein unverzerrter Schätzer der Rate. Für andere Verteilungsmodelle kann man ähnliche Formeln entwickeln. Dieser Umstand legt nahe, ein Schätzverfahren auf der Basis exakter Wartezeiten zu verwenden, das die Nachteile des Sterbetafelschätzers umgeht, ohne spezielle Annahmen über die Verteilung der Wartezeiten vorauszusetzen.

### *4.3.1 Der Ansatz von KAPLAN und MEIER*

Ein solches Verfahren wurde von KAPLAN und MEIER (1958) vorgeschlagen. Die Grundidee ist relativ simpel: Wenn man die Zeitintervalle einer Sterbetafel so klein wählt, daß ein Intervall im Extremfall nur aus einem Zeitpunkt besteht, dann betrachtet man quasi exakte Wartezeiten und

die durch diese Grenzbetrachtung gewonnenen Schätzer (daher auch der Name *Grenzproduktschätzer*, engl. *product–limit–estimator*) haben optimale statistische Eigenschaften. Praktisch verfährt man jedoch etwas anders[1].

In einer Stichprobe werden insgesamt D Ereignisse beobachtet. Seien $t_1 < t_2 < \ldots t_\iota \ldots < t_{D-1} < t_D$ die $\iota = 1,\ldots,D$ Zeitpunkte, zu denen die Ereignisse in der Stichprobe auftreten, dann lassen sich die $i = 1,\ldots,N$ Untersuchungseinheiten an Hand ihrer Ereigniszeitpunkte in einer eindeutigen Reihenfolge anordnen. Das ist dann besonders einfach, wenn jede der N Beobachtungen ein Ereignis zu einem anderen Zeitpunkt hat. In diesem Fall ist der Index $\iota$ mit dem Index i identisch. Häufig werden jedoch mehrere Ereignisse zum gleichen Zeitpunkt beobachtet (ties), so daß die Anzahl I unterschiedlicher Ereigniszeitpunkte nicht mehr der Gesamtzahl D aller Ereignisse entspricht. In diesem Fall ist eine Rangfolge mit einer "Häufung" von Ereignissen zu einzelnen Zeitpunkten möglich.

Schwieriger wird es, wenn einzelne Beobachtungen zensiert sind und gar kein Ereignis aufweisen. Wo sollen sie in der Rangordnung plaziert werden, wenn gar nicht bekannt ist, wann sie ein Ereignis haben? Man verwendet alternativ die Zensierungszeit, und sollten zensierte und unzensierte Beobachtungen zum gleichen Zeitpunkt auftreten, dann vereinbart man, daß die Zensierung etwas später als das Ereignis stattfand. Auf diese Art und Weise ist wieder eine eindeutige Rangordnung möglich, bei der zwar mehrere Ereignisse aber niemals Ereignisse und zensierte Beobachtungen zum gleichen Zeitpunkt stattfinden können. Wenn man jetzt die Zeitachse in hinreichend kleine Zeitintervalle unterteilt, dann treten pro Intervall entweder nur Ereignisse oder nur zensierte Beobachtungen auf und es entfällt die Notwendigkeit, wie beim Sterbetafelschätzer die Risikomenge um einen Anteil zensierter Beobachtungen zu verringern (vgl. Gleichung 4.2). Zum Zeitpunkt des ersten Ereignisses ($\iota = 1$) entspricht die Risikomenge $n_\iota$ dem Stichprobenumfang N. Sie verringert sich mit jedem Intervall um die Anzahl der Ereignisse $d_\iota$ ($d_\iota > 1$ bei ties) oder die Anzahl der zensierten Beobachtungen $c_\iota$. Die Risikomenge $n_\iota$ entspricht praktisch der An-

1) Würde man tatsächlich eine solche Sterbetafel berechnen, wären Sterbetafel– und Kaplan–Meier–Schätzer identisch. Man kann im übrigen zeigen, daß der Kaplan–Meier–Schätzer unter recht allgemeinen Bedingungen ein Maximum–Likelihood–Schätzer von S(t) ist (vgl. LAWLESS 1982: 74ff.); zur Maximum–Likelihood–Methode s. Abschnitt 5.2.

zahl der Beobachtungen, die kurz vor dem $i-$ten Ereigniszeitpunkt noch nicht aus der Untersuchung ausgeschieden sind. Für alle Intervalle, in denen mindestens ein Ereignis auftritt, ist eine Schätzung der (bedingten) Sterbewahrscheinlichkeit möglich:

$$(4.10) \qquad \hat{q}_i = \frac{d_i}{n_i} \quad \text{bzw.} \quad \hat{p}_i = 1 - \hat{q}_i = \frac{n_i - d_i}{n_i}$$

während sie in allen anderen Intervallen (mangels Ereignissen) Null ist:

$$(4.11) \qquad \hat{q}_i = 0 \quad \text{bzw.} \quad \hat{p}_i = 1$$

Die (unbedingte) *Überlebenswahrscheinlichkeit* ergibt sich wiederum durch sukzessive Multiplikation, wobei wegen (4.11) das Produkt nur für die Intervalle berechnet werden muß, in denen Ereignisse auftreten:

$$(4.12) \qquad \hat{S}(t) = \prod_{i|t_i<t} \hat{p}_i = \prod_{i|t<t_i} \frac{n_i - d_i}{n_i}$$

Diese Ausführungen klingen komplizierter als die Berechnung des *Kaplan–Meier–Schätzers* tatsächlich ist. Am besten verdeutlicht man sie sich an einem Beispiel.

Die Daten aus Tabelle 3.2 sind dafür nicht besonders instruktiv, da sehr viele ties auftreten. Die Verwendung des Kaplan–Meier–Schätzers ist also nur dann sinnvoll, wenn man in der Tat über exakte, d.h. kontinuierlich erhobene Wartezeiten verfügt. Ist dies nicht der Fall, dann sollte man gleich den Sterbetafelschätzer verwenden, der für diese Zwecke ein effizienteres und robustes Schätzverfahren darstellt.

Die optimalen Eigenschaften des Kaplan–Meier–Schätzers können also nur dann ausgenutzt werden, wenn die Daten gewissen Genauigkeitsanforderungen genügen. Angenommen, das wäre bei den 10.666 Personen aus Tabelle 3.2 der Fall: Dann entsteht das praktische Problem, wie man diese große Datei in eine eindeutige Reihenfolge bringt. Da jeder Rechnertyp über eigene (maschinenabhängige) Sortierroutinen verfügt, ist das normalerweise kein Problem. Alle sozialwissenschaftlichen Programmpakete, die Kaplan–Meier–Schätzungen anbieten, haben jedoch eigene Sortierroutinen

*Tabelle 4.3: Überlebenswahrscheinlichkeit und empirische kumulierte Rate (KFN – Daten)*

| Daten | | | | Nelson | | Rangordnung | | | | Kaplan – Meier | | Kumulierte Rate | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| i | i* | $t_i$ | $\delta_i$ | $\tilde{H}(t)$ | $\tilde{S}(t)$ | $\iota$ | $t_\iota$ | $d_\iota$ | $n_\iota$ | $\hat{S}(t_\iota)$ | s.e. | $\tilde{H}(t_\iota)$ | $\tilde{S}(t_\iota)$ |
| 1 | 64 | 2 | 1 | 0,0156 | 0,9845 | | | | | | | | |
| 2 | 63 | 2 | 1 | 0,0315 | 0,9690 | | | | | | | | |
| 3 | 62 | 2 | 1 | 0,0476 | 0,9535 | | | | | | | | |
| 4 | 61 | 2 | 1 | 0,0640 | 0,9380 | 1 | 2 | 4 | 64 | 0,9375 | 0,0303 | 0,0625 | 0,9394 |
| 5 | 60 | 3 | 1 | 0,0807 | 0,9225 | | | | | | | | |
| 6 | 59 | 3 | 1 | 0,0976 | 0,9070 | | | | | | | | |
| 7 | 58 | 3 | 1 | 0,1149 | 0,8915 | | | | | | | | |
| 8 | 57 | 3 | 1 | 0,1324 | 0,8760 | | | | | | | | |
| 9 | 56 | 3 | 1 | 0,1503 | 0,8605 | 2 | 3 | 5 | 60 | 0,8594 | 0,0435 | 0,1458 | 0,8643 |
| 10 | 55 | 4 | 1 | 0,1685 | 0,8450 | | | | | | | | |
| 11 | 54 | 4 | 1 | 0,1870 | 0,8295 | 3 | 4 | 2 | 55 | 0,8281 | 0,0472 | 0,1822 | 0,8334 |
| 12 | 53 | 5 | 1 | 0,2058 | 0,8140 | 4 | 5 | 1 | 53 | 0,8125 | 0,0488 | 0,2011 | 0,8179 |
| 13 | 52 | 6 | 1 | 0,2251 | 0,7985 | | | | | | | | |
| 14 | 51 | 6 | 1 | 0,2447 | 0,7830 | 5 | 6 | 2 | 52 | 0,7813 | 0,0517 | 0,2395 | 0,7870 |
| 15 | 50 | 7 | 1 | 0,2647 | 0,7674 | | | | | | | | |
| 16 | 49 | 7 | 1 | 0,2851 | 0,7519 | 6 | 7 | 2 | 50 | 0,7500 | 0,0541 | 0,2795 | 0,7561 |
| 17 | 48 | 7 | 0 | | | | | | | | | | |
| 18 | 47 | 7 | 0 | | | | | | | | | | |
| 19 | 46 | 8 | 0 | | | | | | | | | | |
| 20 | 45 | 8 | 0 | | | | | | | | | | |
| 21 | 44 | 9 | 0 | | | | | | | | | | |
| 22 | 43 | 10 | 1 | 0,3083 | 0,7347 | | | | | | | | |
| 23 | 42 | 10 | 1 | 0,3322 | 0,7174 | 7 | 10 | 2 | 43 | 0,7151 | 0,0570 | 0,3260 | 0,7218 |
| 24 | 41 | 10 | 0 | | | | | | | | | | |
| 25 | 40 | 11 | 0 | | | | | | | | | | |
| 26 | 39 | 12 | 1 | 0,3578 | 0,6992 | | | | | | | | |
| 27 | 38 | 12 | 1 | 0,3841 | 0,6811 | | | | | | | | |
| 28 | 37 | 12 | 1 | 0,4111 | 0,6629 | 8 | 12 | 3 | 39 | 0,6601 | 0,0608 | 0,4029 | 0,6684 |
| 29 | 36 | 14 | 1 | 0,4389 | 0,6447 | | | | | | | | |
| 30 | 35 | 14 | 1 | 0,4675 | 0,6266 | | | | | | | | |
| 31 | 34 | 14 | 1 | 0,4969 | 0,6084 | | | | | | | | |
| 32 | 33 | 14 | 1 | 0,5272 | 0,5903 | 9 | 14 | 4 | 36 | 0,5868 | 0,0641 | 0,5140 | 0,5981 |
| 33 | 32 | 15 | 0 | | | | | | | | | | |
| 34 | 31 | 16 | 1 | 0,5595 | 0,5715 | | | | | | | | |
| 35 | 30 | 16 | 1 | 0,5928 | 0,5528 | | | | | | | | |
| 36 | 29 | 16 | 1 | 0,6273 | 0,5340 | | | | | | | | |
| 37 | 28 | 16 | 1 | 0,6630 | 0,5153 | 10 | 16 | 4 | 31 | 0,5111 | 0,0661 | 0,6431 | 0,5257 |
| 38 | 27 | 16 | 0 | | | | | | | | | | |
| 39 | 26 | 17 | 1 | 0,7015 | 0,4959 | 11 | 17 | 1 | 26 | 0,4914 | 0,0664 | 0,6815 | 0,5058 |
| 40 | 25 | 19 | 0 | | | | | | | | | | |
| 41 | 24 | 20 | 0 | | | | | | | | | | |
| 42 | 23 | 21 | 1 | 0,7449 | 0,4748 | 12 | 21 | 1 | 23 | 0,4700 | 0,0669 | 0,7250 | 0,4843 |
| 43 | 22 | 22 | 0 | | | | | | | | | | |
| : | : | : | 0 | | | | | | | | | | |
| : | : | : | 0 | | | | | | | | | | |
| 64 | 1 | 37 | 0 | | | | | | | | | | |

Zeichenerklärung: Statusvariable $\delta_i$ (1 = Ereignis, 0 = zensierte Beobachtung), s.e. = Standardfehler

Mittelwert der Wartezeiten: t = 22,39 Monate mit Standardfehler s.e. = 1,905 (letzte Beobachtung zensiert bei $t_i$ = 37)

und müssen dafür alle Daten im Programmspeicher halten[1]. Das kann bei großen Dateien zu Platzproblemen führen und beschränkt diesen Ansatz für einige Software-Produkte daher auf kleine und mittlere Datenbestände.

Zur Illustration des Kaplan-Meier-Ansatzes verwende ich an dieser Stelle die Daten über Strafentlassene (KFN-Daten, PRG0402). Alle 64 Probanden sind in Tabelle 4.3 in aufsteigender Reihenfolge je nach Zeitpunkt ihrer Rückfälligkeit bzw. Bewährung aufgeführt. Es handelt sich um einen Prozeß mit einem singulären Ereignis (Rückfall). Alle erfolgreich abgeschlossenen Bewährungszeiten gelten als zensierte Beobachtungen. Die weiteren Berechnungen kann man ebenfalls Tabelle 4.3 entnehmen. Man erkennt, wie die Überlebenswahrscheinlichkeit diskontinuierlich abnimmt. Innerhalb eines Jahres ist nur ein Drittel der Stichprobe rückfällig geworden, nach anderthalb Jahren ist es etwas mehr als die Hälfte. Nach 21 Monaten ist die Überlebenswahrscheinlichkeit nicht mehr definiert, da keine weiteren Ereignisse auftreten.

Bei Verwendung exakter Wartezeiten kann man auch einen Schätzer für den Mittelwert angeben (*durchschnittliche Zustandsdauer*)[2]. Dieser Schätzer ist ein gewichtetes Mittel aller unzensierten Wartezeiten. Für die Gewichtung verwendet man den Kaplan-Meier-Schätzer. Dabei macht man sich folgende Überlegung zunutze: Wenn 100% der Stichprobe die ersten zwei Monate ohne Rückfall übersteht, dann soll diese Zeitstrecke mit dem Faktor 1 gewichtet werden. Wenn 93,8% den nächsten Monat ohne Rückfall übersteht, dann soll diese Zeitstrecke mit dem Faktor 0,938 gewichtet werden usw. Der geschätzte Mittelwert entspricht dann der Summe dieser Zeitstrecken, wobei die Summierung über alle Ereigniszeitpunkte I erfolgt. Wenn die längste Wartezeit aus einer zensierten Beobachtung besteht, dann wird diese Zeitstrecke ersatzweise als letztes "Ereignis" gewertet:

---

1) Dieses Vorgehen ist bei großen Datensätzen, die man besser durch maschinenabhängige Routinen sortiert, nicht besonders effizient. Hier würde man sich wünschen, daß man dem jeweiligen Programmpaket durch einen Befehl mitteilen kann, daß die Daten schon sortiert sind. Das Programm müßte dann nicht alle Daten im Hauptspeicher halten. Bei großen Datensätzen sollte man daher den Sterbetafelansatz verwenden. Angesichts der großen Fallzahl kommen seine negativen Eigenschaften auch nicht zum Tragen.

2) Natürlich kann man auch wieder jedes beliebige Quantil (z.B. den Median) der Verteilung schätzen. Die Berechnung erfolgt analog (4.7).

$$(4.13) \quad \bar{t} = \sum_{\iota}^{l} [\hat{S}(t_{\iota-1})\,(t_\iota - t_{\iota-1})] \quad \text{mit } t_0 = 0 \quad \text{und } \hat{S}(t_0) = 1$$

Solange die längsten Wartezeiten auch mit einem Ereignis enden, ist (4.13) ein akzeptabler Schätzer. Sollten jedoch am Ende der Untersuchung einige Untersuchungseinheiten ohne Ereignis verbleiben, dann unterschätzt (4.13) den wahren Mittelwert. Bei entsprechenden Auswertungen sollte daher in einer Warnung darauf hingewiesen werden, daß auf Grund einer begrenzten Untersuchungsdauer ein Teil der Stichprobe ohne Ereignis verblieb. Angewendet auf die Beispieldaten ergibt sich, daß die untersuchten Personen nach durchschnittlich 22,4 Monaten wieder rückfällig werden. Dabei ist allerdings zu berücksichtigen, daß einige Personen nur maximal 37 Monate beobachtet wurden.

### *4.3.2 Die empirische kumulierte Rate*

Nach Gleichung (2.13) ergibt sich die Überlebenswahrscheinlichkeit S(t) aus der kumulierten Rate H(t). Umgekehrt kann man mit einer Schätzung der Überlebenswahrscheinlichkeit H(t) berechnen:

$$(4.14) \qquad \hat{H}(t) = -\ln \hat{S}(t)$$

Es ist aber auch möglich, die kumulierte Rate direkt mit Hilfe empirischer Daten zu schätzen:

$$(4.15) \qquad \tilde{H}(t) = \sum_{\iota | t_\iota < t} \frac{d_\iota}{n_\iota}$$

Dieser Schätzer wird daher häufig als *empirische (kumulierte) Rate* bezeichnet. Er entspricht quasi der mit der jeweiligen Risikomenge gewichteten Summe der Ereignisse.

Zum besseren Verständnis dieser Schätzgleichung hilft folgende Überlegung: Ausgehend von der zeitlichen Rangordnung der Wartezeiten kann man jeweils die Zeit zwischen zwei Ereignissen $\iota$ und $\iota+1$ messen: $\Delta t_\iota = t_{\iota+1} - t_\iota$. Innerhalb dieser Zeitspanne tritt ein Ereignis (nämlich das $(\iota+1)$-te Ereignis) auf, während der Rest der risikobelasteten Unter-

suchungseinheiten dieses Quasi–Intervall ohne Ereignis übersteht. Geht man wiederum davon aus, daß Ereignisse innerhalb der so definierten Intervalle mit konstanter Rate $\lambda_t$ auftreten, dann führen die gleichen Überlegungen wie in Abschnitt 4.2.2 zu folgendem Schätzer der Rate $\lambda_t$ (vgl. Gleichung 4.8):[1]

$$\tilde{\lambda}_t = \frac{1}{\Delta t_t + (n_t - 1)\,\Delta t_t} = \frac{1}{n_t\,\Delta t_t} \quad \text{für } d_t = 1$$

(4.16)

$$\tilde{\lambda}_t = \frac{d_t}{n_t\,\Delta t_t} \quad \text{für } d_t > 1$$

Der gesamte Prozeß ergibt sich aus der Aneinanderreihung der intervallspezifischen Poisson–Prozesse und die Rate r(t) ist eine Treppenfunktion mit Sprungstellen zu den Zeitpunkten, an denen Ereignisse auftreten. In diesem Fall entspricht S(t):

$$\text{(4.17a)} \qquad S(t) = \exp\left(-\int_0^t r(u)du\right) = \exp\left(-\sum_t \lambda_t\,\Delta t_t\right)$$

und unter Verwendung von (4.16) ergibt sich weiter:

$$\text{(4.17b)} \qquad \tilde{S}(t) = \exp\left[-\sum_{t|t_t<t} \frac{d_t}{n_t}\right] = \exp\left[-\tilde{H}(t)\right]$$

Der Klammerausdruck entspricht der oben definierten empirischen (kumulierten) Rate. Auf diese Art und Weise ergibt sich eine weitere Möglichkeit, die Überlebenswahrscheinlichkeit S(t) zu schätzen.

---

1) Ein Problem stellen wiederum zensierte Beobachtungen dar: Vereinbarungsgemäß sollen sie etwas später stattfinden als das Ereignis zum Zeitpunkt $t$ (s. oben). Für die folgende Ableitung (4.16) ergänzen wir, daß sie nach $t$ aber kurz vor $(t+1)$ ausscheiden. Sie gehen also mit der Intervalldauer $\Delta t_t$ in den Nenner des folgenden Bruches ein. Zur Verdeutlichung betrachten wir die beiden Ereigniszeitpunkte $t_t = 7$ und $t_{t+1} = 10$ mit $\Delta t_t = 3$. Die entsprechenden Summanden für (4.15) lauten 2/50 bzw. 2/43 (5 Zensierungen zwischen $t_t$ und $t_{t+1}$ bei 2/43 ignoriert). Die intervallspezifische Rate $\tilde{\lambda}_t$ entspricht $2/(50 \cdot 3)$, wobei die 5 Zensierungen mit Intervalldauer $\Delta t_t = 3$ berücksichtigt werden.

Dieser zweite Schätzer ist unter der Überschrift "kumulierte Rate" ebenfalls in Tabelle 4.3 aufgeführt und wie man erkennt, unterscheidet er sich nicht wesentlich vom Kaplan–Meier–Schätzer. Das wird auch durch das folgende Argument deutlich. Wenn man in (4.14) die Formel für den Kaplan–Meier–Schätzer einsetzt, erhält man folgenden Ausdruck:

$$(4.18) \qquad \tilde{H}(t) = -\ln\left(\prod_{i|t_i<t} \frac{n_i - d_i}{n_i}\right) = -\sum_{i|t_i<t} \ln\left(1 - \frac{d_i}{n_i}\right)$$

Wenn man den Logarithmus durch eine unendliche Reihe darstellt (vgl. Gleichung A.25), wird deutlich, daß der Schätzer (4.15) quasi eine Approximation erster Ordnung ist[1]. Beide Schätzer unterscheiden sich daher nur für große Werte von t.

Eine weitere Vereinfachung ergibt sich, wenn nur sehr wenige zensierte Beobachtungen und viele unterschiedliche exakte Wartezeiten vorliegen, d.h. eine eindeutige Rangordnung ohne ties existiert. In diesem Fall besteht der Zähler in (4.15) nur aus Einsen und die Risikomenge entspricht der Anzahl der Beobachtungen, die auf das aktuell betrachtete Ereignis folgen plus 1 (das Ereignis selbst). NELSON (1972) hat daher folgende Schätzung der *kumulierten Rate* vorgeschlagen:

1. Zunächst werden alle Wartezeiten in aufsteigender Reihenfolge angeordnet (i = 1, ..., N).
2. Dann werden sie in umgekehrter Reihenfolge durchnumeriert i' = N, ..., 1). Diese Nummern entsprechen genau der jeweiligen Risikomenge.
3. Schließlich berechnet man den Kehrwert der Nummern aus Schritt 2 und summiert diesen bis zur jeweiligen Beobachtung i. Diese Summe ergibt den Schätzer, wenn keine ties existieren (4.15).

Da dieses Verfahren nur eine Sortierung und Numerierung der Daten voraussetzt, ist es praktisch sehr leicht handhabbar. Die entsprechenden Schätzungen finden sich ebenfalls in Tabelle 4.3 unter der Überschrift "Nelson".

1) Formalstatistisch kann man sogar nachweisen, daß der Schätzer (4.15) assymptotisch gleich (4.18) ist (NELSON 1972).

Die dort auftretenden zensierten Beobachtungen werden bei der Summierung (Schritt 3) nicht berücksichtigt.[1]

## *4.4 Gruppenvergleiche*

Nachdem man mit den zuvor besprochenen Methoden einen allgemeinen Überblick über den Veränderungsprozeß bekommen hat, möchte man mögliche Erklärungsfaktoren und Zeitabhängigkeiten untersuchen. Dabei sollte zunächst die erste Teilfrage angegangen werden, da bei mangelnder Kontrolle heterogener Subpopulationen scheinbare Zeitabhängigkeiten entstehen können (vgl. Abschnitt 2.5.1). Dazu wird das Untersuchungsmaterial ähnlich wie in Tabelle 2.6 in $j=1,...,J$ homogene Subgruppen unterteilt. Als Differenzierungskriterium bieten sich alle die Merkmale an, von denen man annimmt, daß sie einen Einfluß auf das Veränderungsgeschehen haben. Durch Vergleich der Funktionen S(t), f(t) und r(t) zwischen den verschiedenen Gruppen erhält man dann erste Aufschlüsse über mögliche Erklärungsfaktoren (Methode des Gruppenvergleichs). Im Rahmen eines Experiments, in dem alle wesentlichen Faktoren durch Randomisierung kontrolliert werden, ist es sogar möglich, Treatment–Effekte durch Vergleich der Experimental– und Kontrollgruppe zu testen. In diesem Fall handelt es sich um eine konfirmatorische (hypothesentestende) Datenanalyse.

Da man nur Subgruppen miteinander vergleichen kann, ist auch klar, daß der Überprüfung erklärender Merkmale enge Grenzen gesetzt sind. So ist es zum Beispiel nicht möglich, metrische Merkmale zu untersuchen, es sei denn, man würde sie vorher klassifizieren. Aus der Kreuztabellenanalyse weiß man darüber hinaus, daß die Anzahl der simultan analysierbaren Variablen wegen der mit der Disaggregierung verbundenen Verringerung der Fallzahl beschränkt ist.

---

1) NELSON's Methode entspricht dem Vorgehen in Gleichung (4.15), wenn man Ereignissen zum gleichen Zeitpunkt nicht denselben sondern einen zufälligen Rangplatz gibt. Bei ties ist ja jede Rangordnung gleich gut. Die Schätzungen nach NELSON's Methode approximieren den Kaplan–Meier–Schätzer sogar noch etwas besser als der Schätzer (4.15).

### *4.4.1 Berechnung von Konfidenzintervallen*

Ich möchte das Vorgehen an Hand der Frage illustrieren, ob sich die Abstiegsrisiken der Berufsanfänger (MZ71 – Daten) mit und ohne Berufsausbildung signifikant unterscheiden. Auf der Basis der Daten in Tabelle 3.8 sind also zwei Sterbetafeln zu berechnen (PRG0403). Wenn man dann die Funktionen S(t), f(t) und r(t) miteinander vergleichen will, ist der Stichprobenfehler zu berücksichtigen. Man braucht daher eine Angabe über die Varianz der verschiedenen Größen, für die jedoch nur in wenigen Fällen eine exakte Ableitung möglich ist, wie sich am Beispiel des Sterbetafelschätzers leicht demonstrieren läßt.

Erinnern wir uns noch einmal an Formel (2.6) für die Überlebenswahrscheinlichkeit. Es handelt sich um nichts anderes als einen Anteilswert, so daß dessen Varianz mit Hilfe der Multinomialverteilung bestimmt werden kann. Für Untersuchungen mit zensierten Beobachtungen ist Formel (2.6) aus den genannten Gründen nicht tauglich, so daß S(t) durch sukzessive Multiplikation bedingter Wahrscheinlichkeiten berechnet werden muß (vgl. Gleichung 4.4). Wiederum können exakte Varianzformeln für die bedingten Wahrscheinlichkeiten angegeben werden, die Varianz des Produktes S(t) ergibt sich jedoch nicht einfach durch Multiplikation. In diesem Fall bedient man sich eines in der Statistik weit verbreiteten Näherungsverfahrens, der sogenannten Delta – Methode[1]. Für den Sterbetafelschätzer von S(t) ergibt sich der folgende Ausdruck, der auch als *Greenwood's Formel* bekannt ist:

$$\hat{\mathrm{VAR}}\,[\hat{S}(\tau_{k-1})] = \hat{S}(\tau_{k-1})^2 \sum_{i=1}^{k-1} \frac{\hat{q}_i}{n_i \hat{p}_i} \tag{4.19}$$

Wenn der Anteil zensierter Beobachtungen pro Intervall groß ist, kann die tatsächliche Varianz beträchtlich unterschätzt werden. Näherungsweise können auch die Varianzen der anderen Verlaufsstatistiken geschätzt werden:

1) Bei der Delta – Methode handelt es sich um ein mathematisches Näherungsverfahren, mit dem nicht – lineare Funktionen durch eine Taylor – Reihe ausgedrückt werden können. Weitere Hinweise findet man bei RAO (1973).

– Varianz der Dichte $f(\tau_{mk})$

$$\hat{f}(\tau_{mk})^2 \left[ \sum_{i=0}^{k-1} \frac{\hat{q}_i}{n_i \hat{p}_i} + \frac{\hat{p}_k}{n_k \hat{q}_k} \right]$$

– Varianz der Rate $r(\tau_{mk})$

$$\hat{r}(\tau_{mk})^2 \left\{ \frac{1 - \left[ \frac{\Delta\tau_k \, \hat{f}(\tau_{mk})}{2} \right]^2}{n_k \, \hat{q}_k} \right\}$$

– Varianz des Medians $\tilde{t}$ der Wartezeiten (j = Median – Intervall)

$$[4 n_j \hat{f}(\tau_{mj})^2]^{-1}$$

– Varianz des Kaplan–Meier–Schätzers für S(t)

$$\hat{S}(t_i)^2 \sum_{i | t_i < t} \frac{d_i}{n_i (n_i - d_i)}$$

– Varianz des Mittelwerts der Wartezeiten $\bar{t}$

$$\sum_{i}^{I} \frac{A_i^2}{n_i \, (n_i - d_i)} \quad \text{mit } A_i = \sum_{j=i}^{I} \hat{S}(t_{j-1}) \, (t_j - t_{j-1})$$

– Varianz der kumulierten Rate H(t)

$$\sum_{i | t_i < t} \frac{d_i}{n_i \, (n_i - d_i)}$$

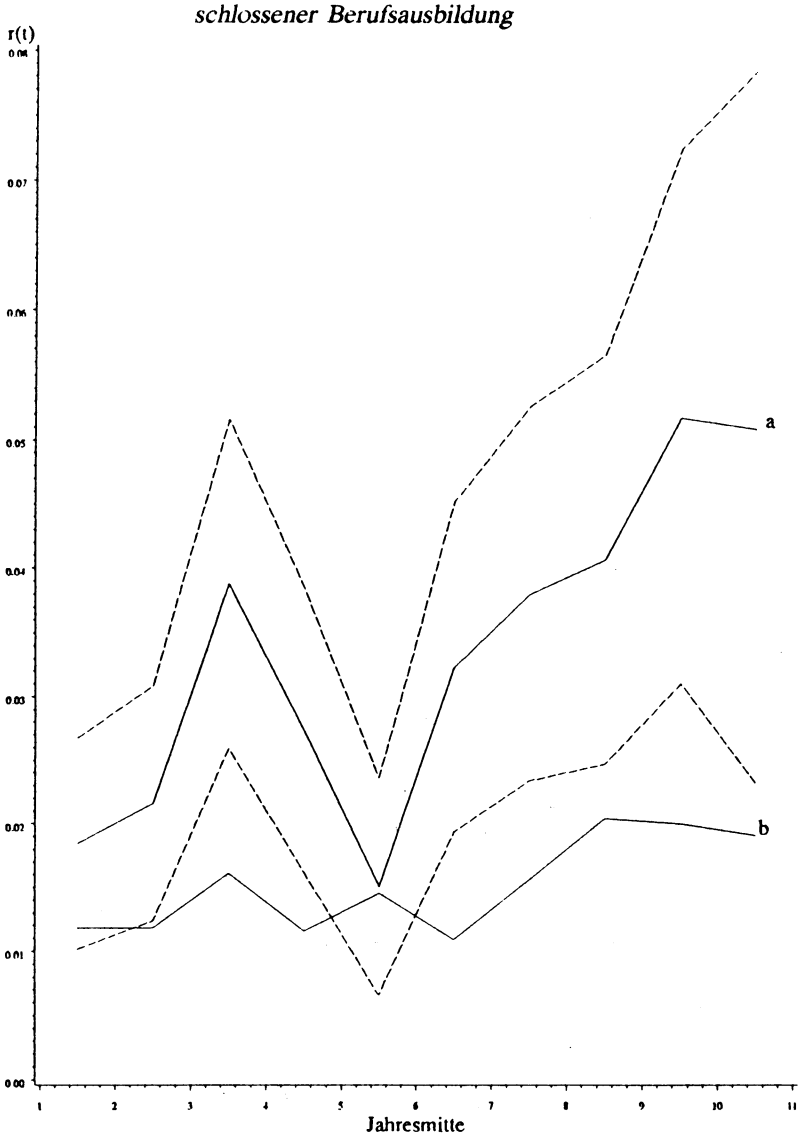
Für alle diese Näherungsformeln gelten ähnliche einschränkende Bemerkungen wie für Greenwood's Ansatz (4.19).

Näherungsweise (d.h. für große Stichproben) sind alle Größen normalverteilt, so daß man mit Hilfe der jeweiligen Varianzformel *Konfidenzintervalle* berechnen kann. Der Bereich beispielsweise, in dem mit 95% Wahrscheinlichkeit der tatsächliche Wert der Rate r(t) liegt, ergibt sich mit Hilfe des geschätzten Standardfehlers $\hat{s}_r$ wie folgt:

$$P[\hat{r}(\tau_{mk}) - 1{,}96\hat{s}_r \leq r(t) \leq \hat{r}(\tau_{mk}) + 1{,}96\hat{s}_r] = 0{,}95 \quad \text{mit } \hat{s}_r = \sqrt{\hat{VAR}[\hat{r}(\tau_{mk})]}$$

Analog kann man Konfidenzintervalle für jeden anderen interessierenden Wert berechnen. Im Hinblick auf meine Ausgangsfrage nach der unterschiedlichen Betroffenheit von beruflichen Abstiegen habe ich die Rate ausgewählt. Abbildung 4.2 zeigt den Verlauf der Abstiegsrate r(t) für

*Abbildung 4.2: Abstiegsrisiko von Personen a) ohne und b) mit abgeschlossener Berufsausbildung*



beide Gruppen. Man erkennt, daß das Abstiegsrisiko der Personen ohne abgeschlossene Berufsausbildung deutlich höher ist als das Abstiegsrisiko der Vergleichsgruppe und zudem im Zeitablauf zunimmt. Auf Grund der geringen Zahl der Abstiege ist das 95%-Konfidenzband der Abstiegsrate (a) relativ groß und überlappt in einigen Intervallen den Verlauf der zweiten Rate (b). In der allgemeinen Tendenz kann man aber von signifikanten Unterschieden in der Abstiegsrate zwischen Personen mit und ohne Berufsausbildung ausgehen. Letztere haben ein signifikantes, zunehmend höheres Abstiegsrisiko im Untersuchungszeitraum.

Strenggenommen sind mit den berechneten Konfidenzintervallen jedoch nur Tests zu einzelnen Zeitpunkten sinnvoll. Aussagen über die gesamte Verteilung sind dagegen auf dieser Basis nicht immer eindeutig, denn es kann der Fall eintreten (wie auch in diesem Beispiel), daß im Intervall k signifikante Unterschiede auftreten, im Intervall k' aber nicht. Daher benötigt man zusätzliche Tests, die die gesamte Verteilung der Wartezeiten berücksichtigen.

### *4.4.2 Nicht-parametrische Rangtests*

Auch bei diesen Testverfahren möchte man zunächst möglichst wenige Annahmen über den zugrundeliegenden Prozeß machen. Solche nicht-parametrischen Tests verwenden in der Regel nur die Rangordnung statt die exakten Zeitpunkte der Ereignisse. Jeder Beobachtung wird dabei ein Rangplatz zugewiesen und die Verteilung der *Ränge* in den einzelnen Subgruppen ist Grundlage der eigentlichen Teststatistik. Dieses Vorgehen hat den Vorteil, daß das Testverfahren relativ robust ist[1]. Es sind darüber hinaus auch keine Verteilungsannahmen notwendig. Natürlich ist mit der Verwendung der Rangordnung ein gewisser Informationsverlust verbunden, jedoch kann gezeigt werden, daß die Effizienz von Rangtests nicht wesentlich geringer ist als die alternativer parametrischer Verfahren. Die Genauigkeit letzterer kommt ohnehin nur dann zur Geltung, wenn ihre Anwendungsvoraussetzungen gegeben sind.

1) Rangtests reagieren weniger sensitiv als parametrische Tests auf sogenannte Ausreißer in den Daten. Jede monotone Transformation der Daten (z.B. logarithmierte Werte statt der Originalwerte) liefert identische Testresultate.

Die statistische Theorie, die hinter diesen Rangtests steht, ist alles andere als trivial. Ausgehend von einem Verteilungsmodell für exakte Wartezeiten überlegt man sich, welche Ränge die einzelnen Beobachtungen erhalten werden. Unterscheiden sich die Verteilungsfunktionen in den einzelnen Subgruppen nur durch ihre zentrale Lage bei ansonsten identischer Dispersion und Form, dann kann man Aussagen über Erwartungswert und Varianz der Rangstatistik machen. Stammen die Daten in der Tat aus einer Grundgesamtheit mit dem unterstellten Verteilungsmodell, dann ist der Rangtest näherungsweise genauso gut wie der entsprechende parametrische Test basierend auf den exakten Wartezeiten. Wenn dagegen das unterstellte Verteilungsmodell in der Grundgesamtheit nicht gilt, dann hat der Rangtest vergleichsweise bessere Eigenschaften. In den Monographien von HAJEK und SIDAK (1967), HAJEK (1969) sowie LEHMANN (1975) findet man weitere Hinweise über das generelle Vorgehen bei Rangtests.

Die hier im Ergebnis zu besprechenden Testverfahren sind im Prinzip Verallgemeinerungen üblicher Rangtests, um zensierte Beobachtungen berücksichtigen zu können. Detailliertere Hinweise findet man bei KALBFLEISCH/ PRENTICE (1980: 143ff.) und LAWLESS (1982: 412ff.). Für den Anwender ist vor allem die Vielzahl der Bezeichnungen für verschiedene Tests verwirrend, die sich nur in Einzelheiten unterscheiden. Am gebräuchlichsten sind die beiden Oberbegriffe *verallgemeinerter Wilcoxon*– und *verallgemeinerter Savage*– oder *Log–Rang–Test.* Häufig werden die Tests aber auch nach ihren jeweiligen "Erfindern" benannt und hier beginnen die Zuordnungsprobleme für den Anwender. Praktisch relevant (d.h. in den meisten Programmpaketen implementiert) sind der verallg. Wilcoxon-Test in der Version von GEHAN (1965) und BRESLOW (1970) sowie der Log–Rang–Test in der Version von COX (1972) und MANTEL (1966).

### 4.4.2.1 Rangziffern

Ausgehend von der (zeitlichen) Rangordnung aller Wartezeiten wird beim verallg. Wilcoxon–Test (in der Version von GEHAN und BRESLOW) jeder Beobachtung i ($i=1,\ldots,N$) ein Wert $\varphi_{ij}$ zugewiesen, der gleich der Zahl der eindeutig kürzeren Wartezeiten abzüglich der Zahl aller eindeutig längeren Wartezeiten ist.[1] Bezogen auf das Anwendungsbeispiel mißt

1) Wir ignorieren zunächst das Auftreten von ties und nehmen an, daß alle Ereignisse jeweils zu einem verschiedenen Zeitpunkt stattfinden. Da jetzt nach verschiedenen Gruppen

daher $\varphi_{ij}$, ob sich mehr Tätigkeitswechsel eindeutig vor oder nach dem Wechsel i ereignen. Negative Werte kennzeichnen sehr kurze Wartezeiten, positive Werte hingegen sehr lange Wartezeiten. Ein Wert von Null beschreibt eine Wartezeit, die genau in der Mitte der Rangordnung steht (Median). Etwas formaler kann man sich überlegen, daß dieser Wert $\varphi_{ij}$ genau gleich der Anzahl der bis dato aufgetretenen Ereignisse (inkl. Ereignis i) entspricht, abzüglich der Risikomenge kurz vor dem Ereignis i:

$$\varphi_{ij} = \sum_{l}^{i} d_{l.} - n_{i.} \tag{4.20}$$

Mit der Formulierung "eindeutig" soll deutlich gemacht werden, daß bei zensierten Beobachtungen nicht genau entschieden werden kann, welche Wartezeit länger dauerte. In diesem Fall trifft man wiederum die Vereinbarung, daß zensierte Beobachtungen immer etwas später auftreten als Ereignisse. Bei zensierten und unzensierten Beobachtungen gleicher Dauer ist also immer die zensierte Wartezeit länger. Das wird in (4.20) durch die Risikomenge berücksichtigt.

Umgekehrt kann für zensierte Beobachtungen nicht entschieden werden, welche Ereignisse eindeutig später auftreten. Wann die jeweilige Untersuchungseinheit ihren Zustand ändert, ist ja auf Grund der Zensierung unbekannt. Daher muß für zensierte Beobachtungen eine spezielle Rangziffer $\phi_{ij}$ eingeführt werden. Sie ist genau gleich der Anzahl der bis dato aufgetretenen Ereignisse, denn nur diese Frage läßt sich eindeutig beantworten:

$$\phi_{ij} = \sum_{l}^{i} d_{l.} \tag{4.21}$$

Beim verallg. Savage–Test werden die Rangziffern nach einer anderen Regel vergeben. Die Rangziffer $\varphi_{ij}$ der i–ten Beobachtung entspricht hier

---

*Forts. von letzter Seite*

differenziert wird, erhält jedes Symbol einen zweiten Index, der die zugehörige Gruppe kennzeichnet. $d_{ij}$ bezeichnet z.B. die Ereignisse in Gruppe j zum Zeitpunkt $t_i$. Ein Punkt bedeutet Summierung über den Index. Alle Ereignisse zum Zeitpunkt $t_i$ werden daher mit $d_{i.}$ bezeichnet.

dem Wert der empirischen kumulierte Rate minus 1. Für zensierte Beobachtungen wird eine Korrektur um +1 vorgenommen:

$$(4.22) \qquad \varphi_{ij} = \sum_{l}^{i} \frac{d_l}{n_l} - 1 \qquad \phi_{ij} = \sum_{l}^{i} \frac{d_l}{n_l}$$

Diese Rangziffern lassen sich weniger anschaulich interpretieren als die zuvor beschriebenen Werte (4.20) und (4.21).

Die eigentliche Teststatistik findet man bei beiden Ansätzen, indem man die Ränge für jede der J Subgruppen aufsummiert. Unter der Nullhypothese, daß alle Subgruppen aus der gleichen Grundgesamtheit stammen, kann man Erwartungswert und Varianz der subgruppenspezifischen Rangsummen berechnen und mit Hilfe der Daten die faktische Identität der Gruppen prüfen.

Bevor ich diese Teststatistik noch auf eine etwas andere Art ableite, möchte ich noch einmal auf die obige Rangordnung eingehen. Aus dem bisher Gesagten wird deutlich, daß das Auftreten *zensierter Beobachtungen* die Berechnung einer eindeutigen Rangordnung erschwert. Die Verallgemeinerung der hier beschriebenen Testverfahren besteht gerade darin, daß sie dennoch diese unvollständigen Beobachtungen berücksichtigen[1]. Das bedeutet aber auch gleichzeitig, daß die Ergebnisse von Rangtests sehr stark von dem Zensierungsmechanismus in den einzelnen Gruppen abhängen. Dieser sollte nicht zu stark zwischen den Gruppen variieren. Um das zu prüfen, untersucht man die Verteilung zensierter Beobachtungen in den einzelnen Subgruppen. *BMDP* erstellt für diese Zwecke eine einfache Graphik[2].

Ein ähnliches Problem entsteht bei sogenannten *ties* (Ereignisse zum gleichen Zeitpunkt). Man kann zwar angeben, welche Wartezeiten eindeutig kürzer bzw. länger sind, aber für die Ereignisse mit identischen Wartezei-

---

1) Wenn keine zensierten Beobachtungen auftreten und die Ränge durch ein logistisches Verteilungsmodell generiert werden, dann ist der verallg. Wilcoxon-Test (in etwas anderer Form) voll effizient. Das gleiche trifft für den verallg. Savage-Test zu, wenn die Ränge durch ein Extremwert-Verteilungsmodell generiert werden.

2) DIEKMANN und MITTER schlagen für diese Zwecke einen einfachen Test vor (1984a: 91): Durch Vertauschen der Ausprägungen "Ereignis" bzw. "Zensierung" kann man mit den hier besprochenen explorativen Methoden prüfen, ob dieses neue "Ereignis" (nämlich die Zensierung) systematisch zwischen den Gruppen variiert.

ten ist praktisch jede Anordnung gleich passend. Mit anderen Worten, man hat in diesem Fall mehrere Möglichkeiten, eine Rangordnung zu vergeben. Sollten relativ zum Stichprobenumfang nicht zu viele ties auftreten, dann kann man diese Variabilität durch entsprechende Varianzformeln berücksichtigen.

### 4.4.2.2 Eine alternative Ableitung

Nach dieser Klarstellung will ich mich eingehender mit der o.g. subgruppenspezifischen Rangsumme beschäftigen. Die jetzt folgende Ableitung der beiden Testverfahren wird deutlich machen, daß sie sich im wesentlichen nur durch die Gewichtung der einzelnen Beobachtungen unterscheiden. $\Omega_j$ sei die Rangsumme für Gruppe j. Sie ergibt sich durch Addition der Rangziffern $\varphi_{ij}$ und $\phi_{ij}$ für alle Beobachtungen innerhalb der Gruppe j. Im Gegensatz zum vorherigen Abschnitt wollen wir nun davon ausgehen, daß mehrere Ereignisse zum gleichen Zeitpunkt auftreten können. Die Summierung kann sich dann auf die $\iota = 1,\ldots,I$ Zeitpunkte ($I \leq D < N$) beschränken, an denen überhaupt Ereignisse auftreten. Sollten zu einem Zeitpunkt mehrere Ereignisse $d_{ij}$ oder zensierte Beobachtungen $c_{ij}$ auftreten, sind die Ränge dementsprechend häufig zu zählen:

$$\Omega_j = \sum_{\iota}^{I} (d_{\iota j}\, \varphi_{\iota j} + c_{\iota j}\, \phi_{\iota j}) = \sum_{\iota}^{I} \omega_{\iota j} \tag{4.23}$$

Durch Einsetzen der entsprechenden Formeln für $\varphi_{ij}$ und $\phi_{ij}$ kann man zeigen, daß sich für den verallg. Savage–Test der Ausdruck (4.24a) und für den verallg. Wilcoxon–Test der Ausdruck (4.24b) ergibt:

$$\Omega_j = -\sum_{\iota}^{I} \left( d_{\iota j} - \frac{d_{\iota\cdot}\, n_{\iota j}}{n_{\iota\cdot}} \right) = -\sum_{\iota}^{I} \psi_{\iota j} \tag{4.24a}$$

$$\Omega_j = -\sum_{\iota}^{I} n_{\iota\cdot} \left( d_{\cdot j} - \frac{d_{\iota\cdot}\, n_{\iota j}}{n_{\iota\cdot}} \right) = -\sum_{\iota}^{I} n_{\iota\cdot}\, \psi_{\iota j} \tag{4.24b}$$

In beiden Fällen mißt der Klammerausdruck $\psi_{ij}$

$$\psi_{ij} = d_{ij} - \frac{d_{i.}\, n_{ij}}{n_{i.}} = d_{ij} - E(d_{ij})$$

die Differenz zwischen tatsächlicher $d_{ij}$ und erwarteter Ereigniszahl $E(d_{ij})$. $E(d_{ij})$ ergibt sich, wenn man die Gesamtzahl der Ereignisse zum Zeitpunkt $i$ proportional zur Risikomenge in den einzelnen Gruppen verteilen würde. Diese Differenz wird dann für alle Zeitpunkte aufsummiert. Beide Testverfahren unterscheiden sich nur durch die Gewichtung dieser Differenzen. Sie werden beim verallg. Wilcoxon-Test mit der jeweiligen Risikomenge multipliziert, während beim verallg. Savage-Test alle Differenzen gleich behandelt werden[1]. Das hat zur Folge, daß der Wilcoxon-Test Unterschiede zu Beginn des Prozesses (Risikomenge noch groß) höher bewertet.

### 4.4.2.3 Ein empirisches Beispiel

Am besten läßt sich das Vorgehen beider Testverfahren an Hand eines Beispiels demonstrieren (PRG0404). An Hand der KFN-Daten[2] möchte ich die Frage prüfen, ob Strafentlassene mit Problemen im sozialen Umfeld ein signifikant höheres Rückfallrisiko im *gesamten* Untersuchungszeitraum haben. Tabelle 4.4 zeigt dazu die nach der Variablen "sozprobl" disaggregierten Informationen aus Tabelle 4.3.

Da der Wilcoxon-Test Unterschiede zu Beginn sehr viel höher bewertet und die Rückfallrisiken zu Beginn noch sehr unterschiedlich sind, ist zu erwarten, daß die Teststatistik des verallg. Wilcoxon-Testes etwas größer ist als die entsprechende Statistik des verallg. Savage-Testes. Tabelle 4.4 enthält sowohl die Berechnung der Rangsumme (4.23) als auch die Addition der Differenzen (4.24) (jeweils für die 1. Gruppe). Beide Summen

---

1) Andere Gewichtungsfaktoren für den verallg. Wilcoxon-Test wurden von PETO/PETO (1972), PRENTICE (1978) sowie TARONE und WARE (1977) vorgeschlagen. PRENTICE verwendet die geschätzte Überlebenswahrscheinlichkeit $\hat{S}(t)$. TARONE und WARE verwenden $\sqrt{n_{i.}}$.

2) Eine Anwendung der Rangtests auf die MZ71-Daten ist prinzipiell möglich, jedoch angesichts der hohen Zahl von ties nicht besonders sinnvoll.

*Tabelle 4.4: Rückfallrisiko und Probleme im sozialen Umfeld (KFN–Daten)*

| | | Gruppe j = 1 mit sozialen Problemen | | | | Gruppe j = 2 ohne soziale Probleme | | | | Insgesamt | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $i$ | $t_i$ | $n_{i1}$ | $d_{i1}$ | $c_{i1}$ | $E(d_{i1})$ | $n_{i2}$ | $d_{i2}$ | $c_{i2}$ | $E(d_{i2})$ | $n_{i.}$ | $d_{i.}$ |
| 1 | 2 | 18 | 4 | 0 | 1,1 | 46 | 0 | 0 | 2,9 | 64 | 4 |
| 2 | 3 | 14 | 5 | 0 | 1,2 | 46 | 0 | 0 | 3,8 | 60 | 5 |
| 3 | 4 | 9 | 1 | 0 | 0,3 | 46 | 1 | 0 | 1,7 | 55 | 2 |
| 4 | 5 | 8 | 0 | 0 | 0,2 | 45 | 1 | 0 | 0,8 | 53 | 1 |
| 5 | 6 | 8 | 0 | 0 | 0,3 | 44 | 2 | 0 | 1,7 | 52 | 2 |
| 6 | 7 | 8 | 1 | 1 | 0,3 | 42 | 1 | 4 | 1,7 | 50 | 2 |
| 7 | 10 | 6 | 0 | 1 | 0,3 | 37 | 2 | 1 | 1,7 | 43 | 2 |
| 8 | 12 | 5 | 1 | 0 | 0,4 | 34 | 2 | 0 | 2,6 | 39 | 3 |
| 9 | 14 | 4 | 2 | 0 | 0,4 | 32 | 2 | 1 | 3,6 | 36 | 4 |
| 10 | 16 | 2 | 1 | 0 | 0,3 | 29 | 3 | 1 | 3,7 | 31 | 4 |
| 11 | 17 | 1 | 0 | 0 | 0,0 | 25 | 1 | 2 | 1,0 | 26 | 1 |
| 12 | 21 | 1 | 0 | 1 | 0,0 | 22 | 1 | 21 | 1,0 | 23 | 1 |

| | verallg. WILCOXON–Test | | | | | verallg. SAVAGE–Test | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $i$ | $\varphi_{i1}$ | $\phi_{i1}$ | $\omega_{i1}$ | $n_i \cdot \psi_{i1}$ | $n_i^2 \cdot \nu_{i11}$ | $\varphi_{i1}$ | $\phi_{i1}$ | $\omega_{i1}$ | $\psi_{i1}$ | $\nu_{i11}$ |
| 1 | −60 | 4 | −240 | 184 | 3154 | −0,938 | 0,063 | −3,750 | 2,875 | 0,770 |
| 2 | −51 | 9 | −255 | 230 | 3002 | −0,854 | 0,146 | −4,271 | 3,833 | 0,834 |
| 3 | −44 | 11 | −44 | 37 | 813 | −0,818 | 0,182 | −0,818 | 0,673 | 0,269 |
| 4 | −41 | 12 | 0 | −8 | 360 | −0,799 | 0,201 | 0,000 | −0,151 | 0,128 |
| 5 | −38 | 14 | 0 | −16 | 690 | −0,760 | 0,240 | 0,000 | −0,308 | 0,255 |
| 6 | −34 | 16 | −18 | 34 | 658 | −0,720 | 0,280 | −0,441 | 0,680 | 0,263 |
| 7 | −25 | 18 | 18 | −12 | 433 | −0,674 | 0,326 | 0,326 | −0,279 | 0,234 |
| 8 | −18 | 21 | −18 | 24 | 483 | −0,597 | 0,403 | −0,597 | 0,615 | 0,318 |
| 9 | −11 | 25 | −22 | 56 | 468 | −0,486 | 0,514 | −0,972 | 1,556 | 0,361 |
| 10 | −2 | 29 | −2 | 23 | 209 | −0,357 | 0,643 | −0,357 | 0,742 | 0,217 |
| 11 | 4 | 30 | 0 | −1 | 25 | −0,318 | 0,682 | 0,000 | −0,038 | 0,037 |
| 12 | 8 | 31 | 31 | −1 | 22 | −0,275 | 0,725 | 0,725 | −0,043 | 0,042 |
| Summe | | | −550 | 550 | 10318 | | | −10,154 | 10,154 | 3,728 |
| Teststatistik | | | | | $X^2 = 29{,}319$ | | | | $X^2 = 27{,}655$ | |

unterscheiden sich, wie zu erwarten, nur durch das Vorzeichen. Varianz und Kovarianz der Summen $\Omega_j$ lassen sich ebenfalls unter den o.g. Annahmen berechnen[1]. Für den verallg. Savage–Test ergibt sich der Ausdruck (4.25a), für den verallg. Wilcoxon–Test der Ausdruck (4.25b):

1) Man erinnere sich daran, daß die Varianzformel bei einem relativ hohen Anteil von ties korrigiert werden muß. Das ist bei (4.25) der Fall.

$$(4.25a)\quad COV(\Omega_j,\Omega_l) = \sum_{i}^{I} \frac{d_{i.}\,(n_{i.} - d_{i.})}{(n_{i.} - 1)} \frac{n_{ij}}{n_{i.}} \left( \delta_{jl} - \frac{n_{il}}{n_{i.}} \right) = \sum_{i}^{I} \nu_{ijl}$$

$$(4.25b)\quad COV(\Omega_j,\Omega_l) = \sum_{i}^{I} n_{i.}^2 \frac{d_{i.}\,(n_{i.} - d_{i.})}{(n_{i.} - 1)} \frac{n_{ij}}{n_{i.}} \left( \delta_{jl} - \frac{n_{il}}{n_{i.}} \right) = \sum_{i}^{I} n_{i.}^2\, \nu_{ijl}$$

$$\text{mit } \delta_{jl} = \begin{cases} 1 & \text{wenn } j = l \quad \text{(Varianz)} \\ 0 & \text{sonst} \quad \text{(Kovarianz)} \end{cases}$$

Unter der Nullhypothese, daß zwei Gruppen aus der gleichen Grundgesamtheit stammen, kann man zeigen, daß das Quadrat der Summe $\Omega_1$ für die erste Gruppe geteilt durch ihre Varianz näherungsweise Chi−Quadrat-verteilt ist (mit einem Freiheitsgrad):

$$(4.26)\qquad X^2 = \frac{\Omega_1^2}{COV(\Omega_1,\Omega_1)} \qquad \text{mit } X^2 \simeq \chi^2_{(1)}$$

Dieses Ergebnis läßt sich leicht auf mehr als zwei Gruppen verallgemeinern und gilt unter recht allgemeinen Bedingungen (vgl. HAJEK/SIDAK 1967: S.159). $\boldsymbol{\Omega}$ sei der Vektor der J subgruppenspezifischen Summen $\Omega_j$ und $\mathbf{V}$ die dazugehörende Varianz−Kovarianz−Matrix mit den Elementen (4.25). Für den Vergleich mehrerer Gruppen gilt dann:

$$(4.27)\qquad X^2 = \boldsymbol{\Omega}'\, \mathbf{V}^{-1}\, \boldsymbol{\Omega} \qquad \text{mit } X^2 \simeq \chi^2_{(j-1)}$$

Da Tabelle 4.4 auch die Varianz (4.25) enthält, kann man sich die Teststatistik für den verallg. Wilcoxon− und Savage−Test berechnen. Wie erwartet ist erstere etwas kleiner. Bei einem Freiheitsgrad sind beide Werte hoch signifikant.

Abschließend kann man über die beiden Rangtests sagen, daß sie unter den o.g. Annahmen (insbes. identische Zensierung) in den verschiedensten Situationen eingesetzt werden können, um die Unterschiede der Überlebensfunktionen verschiedener Subgruppen zu testen. Dabei werden Unterschiede zu Beginn und am Ende der Untersuchung je nach Testverfahren unterschiedlich gewichtet. Beide Tests versagen, wenn sich die subgruppenspezifischen Verläufe von r(t) oder S(t) überschneiden. Dieser Fall kann jedoch

durch eine Inspektion entsprechender graphischer Darstellungen ausgeschlossen werden. Untersuchungen haben darüber hinaus ergeben, daß der verallg. Savage – Test insbesonders dann gut geeignet ist, wenn die Rate in der einen Gruppe ein Vielfaches der Rate in der anderen Gruppe ist (proportionales Risiko), während der verallg. Wilcoxon – Test vor allem dann voll effizient ist, wenn das Verhältnis der Raten in den einzelnen Subgruppen nicht konstant ist. Die beschriebenen Rangtests lassen sich im übrigen so verallgemeinern, daß metrische Kovariate ähnlich wie in der Kovarianz – Analyse kontrolliert werden, während die Unterschiede der Überlebensfunktionen verschiedener Gruppen getestet werden (vgl. KALBFLEISCH/PRENTICE 1980: Kap. 6).

## *4.5 Zeitabhängige Prozesse: Graphische Tests*

Nachdem man das Datenmaterial hinreichend differenziert hat, um möglichst homogene Gruppen zu erhalten, kann man testen, ob es sich um einen zeitabhängigen Prozeß handelt oder nicht. Natürlich lassen sich zeitabhängige Prozesse auch auf der Ebene der Gesamtstichprobe untersuchen, jedoch kann man mögliche Veränderungen im Zeitablauf nicht eindeutig von Einflüssen trennen, die durch heterogene Subpopulationen entstehen. Bevor aber spezielle zeitabhängige Modelle mit parametrischen Methoden getestet werden, sollte man auf jeden Fall eine Exploration der möglichen Modellalternativen durchführen. Man kann dazu *graphische Verfahren* verwenden, denn die Funktionen S(t), f(t) und r(t) müssen einen bestimmten zeitlichen Verlauf aufweisen, wenn ein bestimmtes Verteilungsmodell auf die Daten zutreffen soll. Handelt es sich beispielsweise um einen Prozeß mit exponentiell verteilten Wartezeiten, dann ist die Rate im Zeitablauf konstant und die Überlebenswahrscheinlichkeit eine Exponentialfunktion.

Alle zeitabhängigen Überlebensfunktionen zeigen einen kurvilinearen Verlauf und sind in einigen Fällen (je nach der Wahl der Parameter) visuell kaum voneinander zu unterscheiden. Von daher darf man nicht erwarten, von den hier zu besprechenden Verfahren eindeutige Hinweise auf einen ganz bestimmten Modelltyp zu erhalten. Jedoch lassen sich grobe Modellverstöße eindeutig diagnostizieren und als solche "Notbremsen" sollte man die folgenden graphischen Tests auch verstehen. Da sich kurvilineare Verläufe sehr viel schlechter erkennen lassen als lineare, zeichnet man häufig nicht die Originalwerte von S(t), r(t) etc. sondern geeignete Transformatio-

nen derselben, die einen linearen Verlauf zeigen müssen. Für das o.g. Modell exponentiell verteilter Wartezeiten verwendet man z.B. den natürlichen Logarithmus der Überlebenswahrscheinlichkeit, denn dieser muß wieder eine Gerade ergeben, deren Steigung der (zeitkonstanten) Rate entspricht.

Ganz allgemein ist also eine Transformation $g(\cdot)$ gesucht, die aus der ursprünglich nicht-linearen Funktion für S(t) eine linear-additive verschiedener Zeitfunktionen h(t) macht:

$$g[S(t)] = \alpha_1 h_1(t) + \alpha_2 h_2(t) + \ldots + \alpha_P h_P(t) \tag{4.28}$$

Für die Exponentialverteilung $S(t)=\exp(-\lambda t)$ ist dies, wie gesagt, der natürliche Logarithmus:

$$\ln S(t) = -\lambda t = \alpha_1 t \tag{4.29a}$$

Die Weibull-Verteilung $S(t)=\exp[-(\lambda t)^{\gamma}]$ läßt sich durch die doppelte Logarithmierung linearisieren:

$$\ln[-\ln S(t)] = \gamma \ln\lambda + \gamma \ln t = \alpha_1 + \alpha_2 \ln t \tag{4.29b}$$

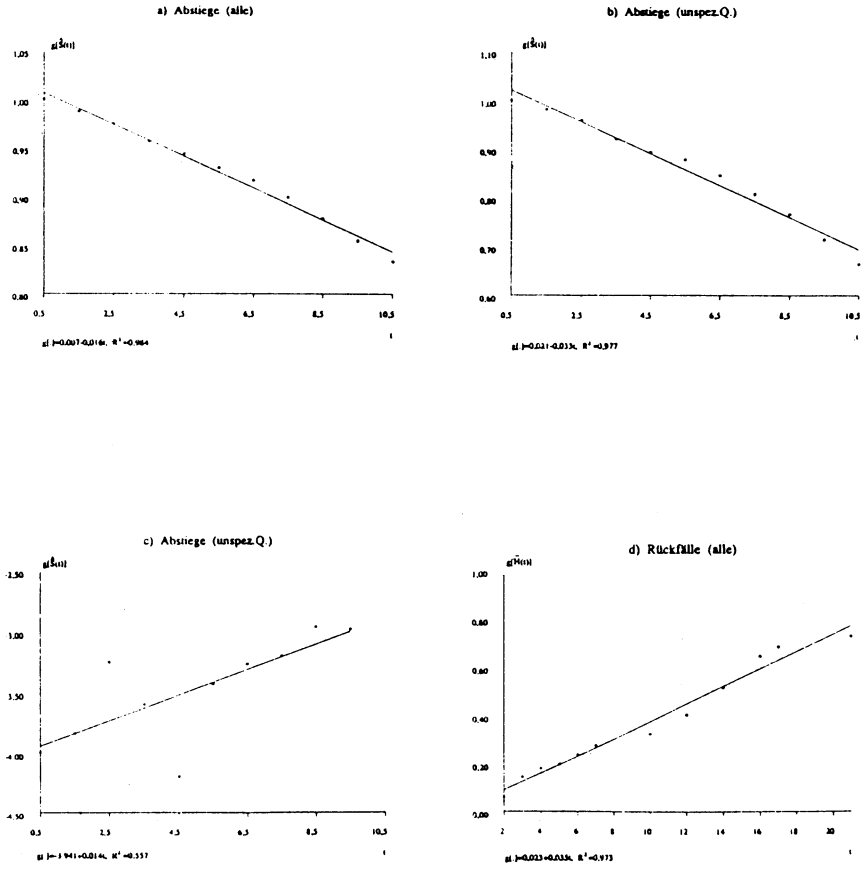
Gompertz-Verteilung $S(t)=\exp[\lambda[1-\exp(\gamma t)]/\gamma]$ und log-logistische Verteilung $S(t)=[1+(\lambda t)^{\gamma}]^{-1}$ können schließlich wie folgt transformiert werden:

$$\ln\left[\ln\frac{S(t_k)}{S(t_{k+1})}\right] = \ln\left[\frac{\lambda}{\gamma}(\exp\gamma - 1)\right] + \gamma t = \alpha_1 + \alpha_2 t \tag{4.29c}$$

$$\ln\frac{1-S(t)}{S(t)} = \gamma\ln\lambda + \gamma\ln t = \alpha_1 + \alpha_2 \ln t \tag{4.29d}$$

Eine Zeichnung der entsprechend transformierten Überlebenswahrscheinlichkeit g[S(t)] (y-Achse) mit t bzw. ln t (x-Achse) sollte daher eine Gerade ergeben, wenn das jeweilige Verteilungsmodell zutrifft. Jedoch ist nicht für alle Verteilungsmodelle des Anhangs B eine einfache Linearisierung möglich (vgl. auch DIEKMANN/MITTER 1984a: 154ff., LAWLESS 1982: 81ff., KALBFLEISCH/PRENTICE 1980: 21ff. und ELANDT-JOHNSON/JOHNSON 1980: 182ff.).

*Abbildung 4.3: Graphische Tests verschiedener Verteilungsmodelle*



Im Prinzip ist es eigentlich egal, welche der drei Funktionen S(t), f(t) oder r(t) gezeichnet wird, denn alle drei Funktionen charakterisieren den gleichen Prozeß (wenn auch auf unterschiedliche Art und Weise). Aus (4.30) ist jedoch ersichtlich, daß man üblicherweise die Überlebenswahrscheinlichkeit S(t) verwendet, da diese Funktion noch am ehesten einen einigermaßen geglätteten Verlauf hat. Der Verlauf der beiden anderen Funktionen r(t) bzw. f(t) ist häufig so unregelmäßig, daß eine eindeutige Zuordnung eines Verteilungsmodells nicht möglich ist. Wenn man schon die Rate verwenden möchte, dann sollte man für die hier zu besprechenden Modelltests die (empirische) kumulierte Risikofunktion verwenden (vgl. Gleichung 4.15).

Mit Hilfe der vorherigen Schätzungen von S(t) bzw. H(t) möchte ich nun verschiedene zeitabhängige Modelle für die MZ71− und KFN−Daten untersuchen. Zunächst einmal interessiert mich die Frage, ob das Abstiegsrisiko in der Tat im Zeitablauf konstant ist, wie in Abschnitt 3.2 vermutet (*exponentiell verteilte Wartezeiten*). Abbildung 4.3a zeigt dazu den natürlichen Logarithmus der Gesamtüberlebenswahrscheinlichkeit, wie er in Tabelle 4.1 geschätzt wurde. Die einzelnen Datenpunkte liegen tatsächlich ungefähr auf einer Geraden, die man entweder frei Hand oder mit einer einfachen Regression bestimmen kann. Im letzten Fall ergibt sich ein Steigungsmaß von −0,016, das man als erste Schätzung für die Abstiegsrate $\lambda$ verwenden kann. Nach Maßgabe des R−Quadrat−Wertes beschreibt die Gerade die Daten sehr gut (98,4% "erklärte" Varianz), jedoch sollte man diese Maßzahl nicht überinterpretieren.

Nach Abbildung 4.2 zu urteilen, sollte das Modell eines zeitkonstanten Abstiegsrisikos nicht für Personen ohne abgeschlossene Berufsausbildung zutreffen. Eine Zeichnung der logarithmierten Überlebenswahrscheinlichkeit in dieser Subgruppe zeigt aber auch einen mehr oder weniger geraden Verlauf und die Regressionsgerade hat einen relativ hohen Datenfit (97,7% "erklärte" Varianz, Abbildung 4.3b). Ich habe dennoch getestet, ob ein *Gompertz−Modell* mit den Daten vereinbar ist. Dafür wurden die (doppelt) logarithmierten Verhältnisse aufeinanderfolgender Schätzungen der Überlebenswahrscheinlichkeit gezeichnet. Das Steigungsmaß der Geraden ist dann ein ungefährer Schätzer des Parameters $\gamma$ (vgl. Gleichung 4.29c). Auf Grund der zunehmenden Risikofunktion in Abbildung 4.2 ist ein positiver Wert zu erwarten.

In der Tat ergibt sich ein Wert von 0,104, jedoch hat die Regressionsgerade einen sehr viel schlechteren Datenfit (55,7% "erklärte" Varianz).

Aus der Abbildung 4.3c wird auch deutlich, worauf dieser mangelnde Fit zurückzuführen ist: Die Werte für das 3. und das 5. Intervall weichen sehr stark von der Geraden ab. Ich würde diese Werte jedoch eher als zufällige Ausreißer einstufen, die die generelle Gültigkeit des Modells nicht einschränken. Auf Grund dieser mehr qualitativen Argumente gibt es zumindest keine Einwände gegen das Gompertz-Modell. Eine Entscheidung für das eine oder andere Modell auf Grund des R-Quadrat-Wertes verbietet sich schon deshalb, weil die Annahmen einer OLS-Schätzung (Homoskedaszität) in diesem Fall nicht gegeben sind. Die solchermaßen berechneten Schätzer für $\lambda$ oder $\gamma$ dürfen daher nur als Anfangswerte für differenziertere Schätzverfahren verwendet werden (vgl. Kapitel 5).

Abschließend habe ich noch untersucht, ob auch das Rückfallrisiko im Zeitablauf konstant ist. Abbildung 4.3d zeigt dazu den Verlauf der empirischen kumulierten Rate, die mit Nelson's Methode geschätzt wurde (vgl. Tabelle 4.3). Nach den Überlegungen in Abschnitt 4.3.2 entspricht sie dem negativen Logarithmus der Überlebenswahrscheinlichkeit. Bei Gültigkeit eines zeitkonstanten Rückfallrisikos erwartet man daher eine ansteigende Gerade. Bis auf einige Unregelmäßigkeiten trifft das auch auf Abbildung 4.3d zu.

Mit diesen Beispielen möchte ich das Thema beenden und noch einmal darauf hinweisen, daß die genannten Verfahren nicht als endgültige Modelltests mißverstanden werden sollten. Für diese Zwecke benötigt man differenziertere Schätzverfahren. Sie sollten nur verwendet werden, um unpassende Modelle auszuschließen (negative Ausgrenzung). Bei der endgültigen Auswahl eines geeigneten Modells für die Daten spielen aber sowohl graphische als auch inhaltliche Argumente eine Rolle. Weitere Hinweise über graphische Verfahren finden sich in dem Buch von NELSON (1982).

## *4.6 Konkurrierende Risiken und wiederholbare Ereignisse*

Alle bisher besprochenen explorativen Verfahren erlaubten nur die Analyse singulärer, nicht-wiederholbarer Ereignisse. Multiple Ereignisse ließen sich nur durch Tricks verarbeiten (Zensierung der konkurrierenden Risiken). Wiederholbare Ereignisse habe ich ganz von der Analyse ausgeschlossen. Es sollte jedoch darauf hingewiesen werden, daß es zumindest zur Analyse konkurrierender Risiken eine recht umfangreiche Literatur gibt

(vgl. etwa LAWLESS 1982: 475ff., KALBFLEISCH/PRENTICE 1980: 163ff., NOMBOODIRI/SUCHINDRAN 1987).

Bei der Berücksichtigung verschiedener Ereignisse mit gruppierten Wartezeiten ergeben sich sogenannte *multiple Sterbetafeln* (vgl. CHIANG 1968, ELANDT-JOHNSON/JOHNSON 1980). In diesem Kontext stellt sich die Frage, ob man bestimmte Risiken statistisch eliminieren kann, um so das eigentlich interessierende Ereignis frei von störenden Randbedingungen untersuchen zu können. Es zeigt sich, daß diese Eliminierung nur unter der schwerwiegenden Annahme unabhängiger Risiken möglich ist. In einer anderen Arbeit (ANDRESS 1984c) habe ich versucht, diesen Ansatz für eine Analyse der Arbeitslosigkeitsdauer fruchtbar zu machen. Ziel war dabei, die Abgänge in Nicht-Erwerbstätigkeit statistisch zu eliminieren und so nur die Zeit bis zur beruflichen Wiedereingliederung zu betrachten. Für diese Berechnungen sind lediglich einige Erweiterungen der oben diskutierten Sterbetafelschätzer notwendig.

Während die Analyse multipler Ereignisse wenn auch nicht besonders komfortabel, so doch zumindest prinzipiell möglich ist, ergeben sich bei Wiederholungen von Ereignissen einige zusätzliche Probleme. Die erste fundamental praktische Frage, die sich in diesem Zusammenhang stellt, lautet: Darf man alle Ereignisse für eine Auswertung zusammenfassen? Darf man also alle beruflichen Abstiege gemeinsam betrachten, obwohl in einer solchen Analyse Personen mit mehreren Abstiegen (d.h. Personen mit spezifischen Eigenschaften) überrepräsentiert sind? Die Antwort hängt wesentlich von den Vorinformationen über den untersuchten Prozeß ab: Wenn die Eigenschaften der Personen mit vielen Ereignissen bekannt sind, kann man die Überrepräsentation der Mobilen durch Disaggregierung in homogene Subgruppen kontrollieren. Des weiteren dürfen keine Einflüsse der Vorgeschichte wirksam sein, denn wenn das der Fall ist, dann müssen auch diese Effekte (Häufigkeit und Dauer früherer Zustände) kontrolliert werden.

Am einfachsten erkennt man vielleicht das Problem, wenn man sich überlegt, bei welchen Prozessen ohne weiteres *wiederholbare Ereignisse* betrachtet werden können. Angenommen es handelt sich um eine homogene Population, in der Ereignisse mit konstanter Rate auftreten (Poisson-Prozeß). Hier macht es in der Tat keinen Unterschied, ob man alle oder nur die jeweils ersten Wartezeiten betrachtet. Das gleiche trifft für einen Prozeß zu, in dem in einer homogenen Population je nach Dauer des aktuellen Zustands Ereignisse unterschiedlich wahrscheinlich sind (Semi-Markov-

Prozeß). Da diese Abhängigkeit von der Zustandsdauer für jeden Zustand gleich ist, macht es keinen Unterschied, ob man alle oder nur die jeweils ersten Ereignisse betrachtet. Probleme entstehen nur dann, wenn heterogene Subpopulationen oder Einflüsse der Vorgeschichte auftreten. In diesem Fall entsteht eine ähnliche Situation wie bei der oben beschriebenen Analyse zeitabhängiger Prozesse ohne ausreichende Kontrolle der Heterogenität des Datenmaterials. Welche Fehlschlüsse sich jedoch bei der Analyse wiederholbarer Ereignisse ergeben können, ist nicht einfach anzugeben. Man sollte daher generell sehr vorsichtig sein, wenn man bei den hier besprochenen explorativen Methoden alle Ereignisse gemeinsam betrachtet. Auf jeden Fall sollte durch getrennte Analysen der jeweils ersten und aller Ereignisse geprüft werden, ob die Berücksichtigung von Ereigniswiederholungen wesentlich andere Ergebnisse liefert.

# 5. *Konfirmatorische Verfahren*

Nachdem man durch geeignete Methoden einen ersten Überblick über die Daten erhalten hat, geht es nun darum, spezifische Fragestellungen an Hand des empirischen Materials zu überprüfen. In der Regel sollen dabei komplexe Hypothesen mit mehreren Erklärungsfaktoren untersucht werden, so daß die einfachen Methoden des vorigen Kapitels nicht mehr ausreichen. In diesem Fall benötigt man Modelle, die es gestatten, den beobachteten Prozeß auf Kovariate und zeitliche Veränderungen zurückzuführen, und die dabei gleichzeitig den gemeinsamen Einfluß der verschiedenen Erklärungsfaktoren berücksichtigen.

Im folgenden Abschnitt 5.1 möchte ich die Eigenschaften eines Erklärungsmodells für Verlaufsanalysen diskutieren. Dazu sollte man sich an die wesentlichen Auswertungsschritte bei einem linearen Regressionsmodell erinnern (vgl. z.B. HANUSHEK/JACKSON 1977 oder DRAPER/SMITH 1981). Ein Vergleich mit diesem klassischen Ansatz sozialwissenschaftlicher Datenanalyse verdeutlicht am besten, wo die speziellen Probleme bei der Analyse von Verlaufsdaten liegen. Auf der anderen Seite zeigt sich auch, daß Regressionsmodelle für Verlaufsdaten sehr viele Gemeinsamkeiten mit den klassischen Regressionsmodellen haben. Entgegen allen Vorurteilen ist es daher nur an einigen Stellen notwendig, bisher bekannte Auswertungsstrategien abzuändern und zu erweitern. Am Ende dieses Vergleichs (Abschnitt 5.1.5) findet sich eine kommentierte Inhaltsangabe des folgenden Kapitels.

## *5.1 Regressionsmodelle für Verlaufsdaten — Überblick*

### *5.1.1 Multiple Regression — Eine Wiederholung*

Die wesentlichen Schritte bei der Spezifikation und Überprüfung eines klassischen Regressionsmodells lassen sich wie folgt zusammenfassen (vgl. auch Abb. 3.4):

1. Modellspezifikation,
2. Schätzung der Modellparameter,
3. Auswertung und Modelltests,

4. Interpretation der Ergebnisse und
5. Modellevaluation (Residuenanalyse, Annahmentests).

Angenommen man möchte Einkommensunterschiede untersuchen, dann muß man zunächst einmal festlegen, welche sozio-demographischen Merkmale in welcher Weise die Zielvariable Einkommen bestimmen sollen (Modellspezifikation). Man bevorzugt hierbei linear-additive Zusammenhänge, obwohl diese Wahl nicht zwingend vorgegeben ist. Bei einem solchen Modell hat man natürlich nicht den Anspruch, alle Einkommensunterschiede erklären zu können. Daher enthält die Modellgleichung zusätzlich einen Fehlerterm. Etwas formaler ausgedrückt wird die Zielvariable y in eine *systematische Komponente* $\mu$ und eine *Fehlerkomponente* $\epsilon$ zerlegt, wobei sich das Erklärungsmodell auf $\mu$ oder den Erwartungswert von y bezieht.

$$(5.1) \quad y_i = \mu_i + \epsilon \quad \text{mit } \mu_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \ldots + \alpha_P x_{iP}$$

Wie angegeben ist der Erwartungswert eine linear-additive Funktion der $p=1,\ldots,P$ Kovariaten $x_{ip}$. Sei $\mathbf{x}_i$ der Zeilenvektor dieser Kovariaten $[x_{i0},x_{i1},\ldots,x_{ip}]$ erweitert um eine Variable $x_{i0}$, die für alle Beobachtungen den Wert 1 hat, und $\boldsymbol{\alpha}$ der entsprechende $(P+1)\cdot 1$-Spaltenvektor der Regressionskoeffizienten, dann lautet das Regressionsmodell: $y_i = \mathbf{x}_i\boldsymbol{\alpha} + \epsilon$.

Der nächste Schritt besteht nun darin, die unbekannten Parameter des Modells mit Hilfe empirischer Daten zu schätzen. Die Parameter werden dabei so gewählt, daß die Modellprognosen (d.h. die unter den Modellannahmen erwarteten Werte von y) minimal von den empirischen Werten abweichen. Genauer gesagt werden die quadrierten Abweichungen minimiert und dieses Schätzverfahren hat daher den Namen *Kleinste-Quadrate-Schätzung* (LS – *L*east *S*quares). Unter bestimmten Voraussetzungen haben die gefundenen Schätzer optimale statistische Eigenschaften. Kann man darüber hinaus annehmen, daß die Fehler $\epsilon$ normalverteilt sind, dann lassen sich die üblichen Signifikanztests (t- und F-Test) verwenden, um einzelne Parameter oder den gesamten Modellfit zu überprüfen.

In Schritt 3 geht es um eine Auswertung der Schätzergebnisse im Hinblick auf die Ausgangshypothese(n). Dabei spielen folgende Fragen eine Rolle:

- Gibt das Modell eine angemessene Beschreibung der Daten (Modellfit)?
- Ist es signifikant besser als andere (möglicherweise einfachere) Modelle?

- Zeigen die geschätzten Parameter die erwartete Richtung und Größe?
- Unterscheiden sie sich signifikant von den Werten, die man ganz bestimmt ausschließen möchte (z.B. dem Wert Null)?

Von der Beantwortung dieser Fragen hängt ab, ob man das gesamte Modell als passend akzeptiert.

Ist das der Fall, dann kann man dazu übergehen, die Modellergebnisse in möglichst anschaulicher Form zu interpretieren (Schritt 4). Wird dagegen das Modell verworfen, dann beginnt der ganze Prozeß von vorne. Bei dieser Abänderung und Respezifikation ist es hilfreich zu wissen, wo das ursprüngliche Modell versagt hat. Man wird sich daher die Einkommenswerte genauer anschauen, die ganz besonders schlecht vorhergesagt wurden (große Residuen): Durch welche Merkmale lassen sie sich kennzeichnen und können diese Variablen in einem neuen Modell berücksichtigt werden? Auch wird man die Modellannahmen näher untersuchen: Ist es z.B. richtig, daß das Einkommen linear mit der Qualifikation einer Person zunimmt? Alle diese weitergehenden Analysen fasse ich unter dem Oberbegriff Modellevaluation zusammen. Idealiter steht am Ende dieses Schrittes 5 eine realistischere Vorstellung über die Daten, die es gestattet, ein verbessertes Modell zu formulieren, dessen Parameter dann erneut mit Hilfe der Daten geschätzt werden können.

Modelle sind, wie der Name sagt, nur ein Abbild der Realität unter gewissen vereinfachenden Annahmen. Zu einer seriösen Information über Ergebnisse statistischer Methoden gehört daher auch die Aufklärung über die Annahmen des Modells. Man beachte, daß die Strategien, die ich unter dem Oberbegriff Modellevaluation zusammengefaßt habe, nichts anderes sind als ein Test der expliziten (z.B. linear–additives Modell) und impliziten Modellannahmen (z.B. aus dem Modell resultierende Prognosen). Schritt 5 versteht sich daher nicht nur als ein Zwischenschritt auf dem Weg zum passenden Modell, sondern auch als ein wesentlicher Bestandteil der Interpretation der Modellergebnisse. Im übrigen ist eine anschauliche Interpretation gerade mit Modellprognosen sehr gut möglich.

### *5.1.2 Regressionsmodelle für Verlaufsdaten – Raten als Zielvariable*

Betrachten wir nun Regressionsmodelle für Verlaufsdaten, wobei wir mit der am häufigsten verwendeten Variante, dem Modell proportionaler Risiken, beginnen. Der auffallendste Unterschied zu dem vorherigen Regres-

sionsmodell für Einkommensunterschiede ist die Tatsache, daß die Zielvariable jetzt aus einer nicht-beobachtbaren Größe (der Rate) besteht, die erst durch empirische Daten geschätzt werden muß. Dieser Umstand hat zwei schwerwiegende Konsequenzen, wie man gleich sehen wird. Beschäftigen wir uns jedoch zunächst mit den Annahmen dieses Modells.

#### 5.1.2.1 Schritt 1: Modellspezifikation

Zu diesem Zweck verwende ich eine etwas allgemeinere Formulierung, aus der die Eigenschaften des Modells besser deutlich werden.

$$r(t|\mathbf{x}_i) = \lambda_0(t)g(\mathbf{x}_i) \tag{5.2a}$$

In dieser Gleichung wird die funktionale Abhängigkeit $g(\cdot)$ der Rate von den $p=1,...P$ Kovariaten $\mathbf{x}_i$ zunächst nicht spezifiziert. $\lambda_0(t)$ ist eine für alle Untersuchungseinheiten identische *Basisrate*, die Veränderungen des Prozesses im Zeitablauf beschreibt. Dieses Modell bezeichnet man als ein *Modell proportionaler Risiken* (engl. proportional hazards model), da sich die Risiken zweier Untersuchungseinheiten i und j zu allen Zeitpunkten nur durch einen konstanten Proportionalitätsfaktor c unterscheiden. Die Rate $r_i(t)$ ist also immer ein konstantes Vielfaches der Rate $r_j(t)$:

$$\frac{r_i(t|\mathbf{x}_i)}{r_j(t|\mathbf{x}_j)} = \frac{\lambda_0(t)g(\mathbf{x}_i)}{\lambda_0(t)g(\mathbf{x}_j)} = \frac{g(\mathbf{x}_i)}{g(\mathbf{x}_j)} = c \tag{5.3}$$

Diese Beziehung gilt unabhängig davon, welchen Wert die Funktion $\lambda_0(t)$ aufweist, vorausgesetzt die Basisrate verändert sich für alle Untersuchungseinheiten in gleicher Weise. Man beachte, daß diese Gleichung für das folgende Modell nicht gilt:

$$r(t|\mathbf{x}_i) = \lambda_0(t) + g(\mathbf{x}_i)$$

Hier wäre nur die Differenz der Raten konstant.

Man beachte außerdem, daß zeitliche Veränderungen des Prozesses von den Kovariaten unabhängig sind. Oder anders ausgedrückt: Es besteht in diesem Modell keine Möglichkeit, ein zu- oder abnehmendes Risiko zu modellieren, dessen Veränderung mit bestimmten Merkmalen zusammenhängt. Diese Unabhängigkeit von zeitlicher Entwicklung und Kovariaten

ergibt des weiteren eine besonders einfache Formel für die Überlebenswahrscheinlichkeit. Unter Verwendung von (2.13) erhält man nämlich:

$$(5.4a) \qquad S(t|\mathbf{x}_i) = \exp[-\int_0^t \lambda_0(u)g(\mathbf{x}_i)\,du] = [S_0(t)]^{g(\mathbf{x}_i)}$$

$$\text{mit } S_0(t) = \exp[-\int_0^t \lambda_0(u)du]$$

Mit anderen Worten, die allgemeine Überlebenswahrscheinlichkeit für beliebige Konstellationen der Kovariaten erhält man durch eine Potenz der *Basis–Überlebenswahrscheinlichkeit* $S_0(t)$, die allein eine Funktion der *Basisrate* $\lambda_0(t)$ ist.

Implizit wird damit gefordert, daß sich die Überlebensfunktionen nicht überkreuzen dürfen. Diese Annahme läßt sich durch eine Zeichnung der Überlebensfunktionen für verschiedene Subgruppen leicht überprüfen. Dabei spielt keine Rolle, wie die Kovariaten wirken, wie also $g(\mathbf{x}_i)$ konkret aussieht (log–linear oder linear–additiv). Sollten sich die Überlebensfunktionen überkreuzen, dann ergibt sich mit (5.4a) auch gleich eine Verallgemeinerung des Modells. In diesem Fall fordert man, daß Gleichung (5.4a) nur innerhalb einer Subgruppe gilt, die Basisrate also für jede Subgruppe verschieden ist.

Als nächstes möchte ich die funktionale Abhängigkeit von den Kovariaten betrachten. Da die Rate r(t) keine negativen Werte annehmen kann, bieten sich log–lineare Abhängigkeiten $g(\mathbf{x}_i)=\exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_P x_{iP})$ an. Sei $\mathbf{x}_i$ der $1 \cdot P$–Zeilenvektor dieser $p=1,\ldots,P$ Kovariaten und $\boldsymbol{\beta}$ der entsprechende $P \cdot 1$–Spaltenvektor der Regressionskoeffizienten, dann ergibt sich folgendes log–lineares Regressionsmodell proportionaler Risiken:[1]

$$(5.2b) \qquad r(t|\mathbf{x}_i) = \lambda_0(t)\exp(\mathbf{x}_i\boldsymbol{\beta})$$

Für die Überlebenswahrscheinlichkeit (5.4a) gilt jetzt:

$$(5.4b) \qquad S(t|\mathbf{x}_i) = [S_0(t)]^{\exp(\mathbf{x}_i\boldsymbol{\beta})}$$

---

1) Man beachte das Fehlen einer Regressionskonstanten, die ggfs. in dem Term $\lambda_0(t)$ unterzubringen ist.

Dabei ist es hilfreich eine doppelt logarithmierte Transformation von S(t) zu betrachten:

$$\ln[-\ln S(t|\mathbf{x}_i)] = \mathbf{x}_i\beta + \ln[-\ln S_0(t)] \tag{5.5}$$

Aus dieser Umformung erkennt man, daß die doppelt logarithmierten Überlebensfunktionen zweier Gruppen im konstanten Abstand zueinander verlaufen müssen, wenn sich beide Gruppen nur durch einen konstanten Faktor $\mathbf{x}_i\beta$ unterscheiden.

*Log–lineare Abhängigkeiten* sind nicht so einfach zu interpretieren. Linearität und Additivität sind jetzt nur noch bei Betrachtung der logarithmierten Raten gegeben:

$$\ln r(t|\mathbf{x}_i) = \ln \lambda_0(t) + \mathbf{x}_i\beta$$

Will man aber Aussagen über die Rate machen, dann muß der Einfluß der Kovariaten multipliziert werden. Diese Eigenschaft des Modells erweist sich insbesondere bei der anschaulichen Interpretation der Ergebnisse als sehr störend.

Natürlich sind auch andere funktionale Abhängigkeiten denkbar, z.B. *linear–additive*. Hier kann jedoch nicht ausgeschlossen werden, daß das Modell negative Werte vorhersagt. Eine negative Rate wäre jedoch eine sinnlose Prognose. Unabhängig von dieser unschönen Eigenschaft würde aber noch die Annahme der Proportionalität (5.3) gelten. Allerdings läßt sich Gleichung (5.2b) nicht mehr so einfach schreiben: Jeder Parameter müßte quasi mit der Basisrate multipliziert werden.

#### 5.1.2.2 Schritt 2: Schätzung der Modellparameter

Das Modell proportionaler Risiken unterscheidet sich also schon bei der Modellspezifikation (Schritt 1) von bisher bekannten Abhängigkeitsbeziehungen bei Regressionsmodellen. Praktische Probleme werden jedoch die meisten Anwender bei der Frage haben, wie man Modell (5.2) mit Hilfe empirischer Daten überprüft, da hier nicht–beobachtbare (die Rate) mit beobachtbaren Größen (den Kovariaten und der Zeit) in Verbindung gebracht werden. Das Schätzverfahren ist jedoch in seinen Grundprinzipien ähnlich simpel wie bei LS–Schätzungen.

Die Grundüberlegung läßt sich in vereinfachter Form wie folgt darstellen: Wenn man einen empirischen Prozeß erhebt, über den man gewisse theoretische Vorstellungen hat (z.B. Ereignisse treten mit konstanter Rate auf), dann kann man unter den Modellannahmen Aussagen über die Wahrscheinlichkeit einzelner Beobachtungen machen. Z.B. über die Frage, wie wahrscheinlich ein Ereignis innerhalb eines Zeitintervalls $\Delta t$ ist. Wie man weiß, hängt diese Wahrscheinlichkeit von der Rate ab und kann daher unter den Modellannahmen (5.2) auf die Kovariaten zurückgeführt werden. Durch Rückgriff auf die Wahrscheinlichkeit der Stichprobenbeobachtungen ist es also möglich, empirische Größen (Ereignisse, Zensierungen) mit den erklärenden Variablen des Modells in Verbindung zu bringen, obwohl in (5.2) eigentlich eine nicht beobachtbare Variable modelliert wird. Für die untersuchte Zufallsvariable kann man daher sagen: Die Parameter des Modells werden so gewählt, daß die Gesamtwahrscheinlichkeit aller Stichprobenbeobachtungen maximal wird[1]. Mit anderen Worten, die Parameter werden so geschätzt, daß die vorliegende Stichprobe unter den Modellannahmen die plausibelste ist.

Das Schätzverfahren heißt Methode der größten Mutmaßlichkeit oder *ML – Schätzung* (**M**aximum Likelihood). Es setzt voraus, daß man ein Modell für die Daten angeben kann, aus dem die Wahrscheinlichkeitsdichte jeder Beobachtung abgeleitet werden kann. Etwas formaler ausgedrückt, es muß möglich sein, eine Likelihoodfunktion für die Stichprobe zu benennen. Das ist dann besonders schwierig, wenn nicht alle Teile des Modells spezifiziert sind. In dem *Regressionsmodell von COX* (1972), das ein Spezialfall des allgemeinen Modells proportionaler Risiken ist, wird z.B. keine Aussage über die Art der Zeitabhängigkeit $\lambda_0(t)$ gemacht. Man nennt dieses Modell daher auch *partiell parametrisch*. In diesem Fall ist es notwendig, Abwandlungen des ML – Schätzverfahrens zu verwenden, die unter dem Namen *Partial Likelihood* bekannt geworden sind.

Erste Konsequenz der nicht – beobachtbaren Zielvariablen ist also ein anderes Schätzverfahren. Eine zweite, weniger offensichtliche Konsequenz ist das Fehlen eines Fehlerterms in der Modellgleichung. Da hier Wahrscheinlichkeiten bzw. Dichten modelliert werden, ist das Modell quasi von Anfang an stochastisch und man hat es daher nicht für nötig gefunden,

1) Strenggenommen gilt diese Formulierung nur für diskrete Zufallsvariablen. Bei kontinuierlichen Zufallsvariablen betrachtet man die Dichte. Daher auch die Bezeichnung "Likelihood" statt "Probability".

einen Fehlerterm zu berücksichtigen. Es kann auch gezeigt werden, daß die resultierenden ML-Schätzer die besten sind, vorausgesetzt das Modell ist richtig. Die Konsequenzen einer Fehlspezifikation sind jedoch nur teilweise bekannt und in einigen Fällen hat es sich daher als sinnvoll erwiesen, einen Fehlerterm zu berücksichtigen.

ML-Schätzungen haben gegenüber LS-Schätzungen einen weiteren Vorteil, wenn man berücksichtigt, daß ein Teil der Verlaufsdaten immer unvollständig ist. In der Likelihood-Funktion kann man nämlich unterscheiden, ob es sich bei einer Beobachtung um ein Ereignis oder um eine zensierte Beobachtung handelt. Würde man hingegen wie in einer LS-Schätzung Abweichungen von einer beobachtbaren Größe (z.b der Wartezeit) minimieren, dann stünde man vor der Schwierigkeit, daß nur für einen Teil der Beobachtungen vollständige Informationen vorliegen. Auch wenn ML-Schätzungen auf den ersten Blick sehr viel komplizierter ausschauen, erlauben sie also, sehr viel differenzierter mit den Daten umzugehen.

#### 5.1.2.3 Schritt 3-5: Auswertung, Interpretation und Evaluation der Ergebnisse

Nachdem man die Parameter des Modells geschätzt hat, folgen die Schritte 3-5 analog dem Vorgehen bei klassischen Regressionsmodellen. Auch bei ML-Schätzungen kann man unter gewissen allgemeinen Bedingungen die üblichen Signifikanztests verwenden, um Hypothesen und Modellfit inferenzstatistisch zu prüfen. Dagegen ist die anschauliche Interpretation der Modellergebnisse sowie eine deskriptive Beschreibung des Modellfits nicht so einfach.

Das erste Problem hat etwas damit zu tun, daß sich die unterstellten *log-linearen Abhängigkeiten* nicht so leicht interpretieren lassen wie linear-additive. Eine anschauliche Interpretation log-linearer Parameter ist vor allem dann möglich, wenn es gelingt, die Parameterschätzungen auf beobachtbare Größen (z.B. die durchschnittliche Verweildauer in einem Zustand) zurückzubeziehen.

Das zweite Problem hat wieder etwas damit zu tun, daß ein Teil der Beobachtungen unvollständig ist. In diesem Fall macht es keinen Sinn, eine Maßzahl wie den *Anteil erklärter Varianz* zu berechnen, denn für einen Teil der Beobachtungen ist die Berechnung eines Residuums gar nicht

möglich, da schon die Angabe über die abgeschlossene Wartezeit fehlt[1]. Aus dem gleichen Grunde kann man bei Verlaufsdaten nicht so ohne weiteres eine *Residuenanalyse* durchführen.

Nach dieser Aufzählung der wesentlichen Unterschiede und Probleme ist es eigentlich an der Zeit, darauf hinzuweisen, wo die aufgeworfenen Fragen in diesem Kapitel (hoffentlich) beantwortet werden. Zuvor möchte ich aber auf einige Ähnlichkeiten zwischen klassischen Regressionsmodellen für metrische Daten und den hier diskutierten Modellen für Verlaufsdaten hinweisen. Konkret möchte ich zeigen, daß das erwähnte Modell proportionaler Risiken sich auch wie ein klassisches Regressionsmodell schreiben läßt, in dem eine beobachtbare Größe die Zielvariable ist.

### *5.1.3 Regressionsmodelle für Verlaufsdaten — Wartezeiten als Zielvariable*

Das o.g. Regressionsmodell zur Erklärung von Einkommensunterschieden kann auch auf etwas andere Art und Weise eingeführt werden: Dazu sei angenommen, daß Einkommensschwankungen *normalverteilt* sind. Etwas formaler ausgedrückt bedeutet das, daß die standardisierten Abweichungen vom Erwartungswert standardnormalverteilt sind:

$$z_i = \frac{y_i - \mu}{\sigma} \qquad \text{mit } Z \sim N(0,1)$$

Mit dieser Annahme kann man bekanntermaßen die Wahrscheinlichkeit bestimmter Einkommensgruppen berechnen. Nimmt man weiter an, daß das erwartete Einkommen $\mu$ mit den Eigenschaften der untersuchten Person variiert ($\mu_i = \mathbf{x}_i\alpha_i$), dann ergibt eine Umformung das Regressionsmodell (5.1), das jeden Einkommenswert $y_i$ in eine systematische Komponente $\mu_i$ und in eine Fehlerkomponente $\epsilon$ zerlegt:

$$y_i = \mu_i + \epsilon \qquad \text{mit } \epsilon = \sigma z \quad \text{und } Z \sim N(0,1)$$

---

1) Man beachte, daß das Residuum im ML–Algorithmus nicht verwendet wird, während es bei LS–Schätzungen eine zentrale Rolle spielt: Dort wird ja bekanntlich die Quadratsumme der Residuen minimiert.

Man erkennt, daß der Fehlerterm aus einer standardnormalverteilten Zufallsvariablen Z mit konstanter Varianz $\sigma$ besteht. Während ich in Abschnitt 5.1.1 zuerst ein Prognosemodell und erst dann eine Verteilungsannahme über den Fehlerterm gemacht habe, ergibt sich hier das Prognosemodell und der Fehlerterm aus einer globalen Verteilungsannahme über die Zielvariable. Die Normalverteilung hat dabei die schöne Eigenschaft, nur zwei Parameter zu besitzen, wovon der eine $\mu$ die *zentrale Lage* (location) und der andere $\sigma$ die *Dispersion* (scale) der Verteilung beschreibt. Das Regressionsmodell macht praktisch nur Aussagen über den Lageparamter der Verteilung. LAWLESS (1982) bezeichnet diese Modelle daher auch als "location−scale models". Wenn es nun gelänge, ähnliche "location−scale models" für zeitbezogene Daten zu finden, dann hätte das den Vorteil, daß man ähnlich (5.1) lineare Regressionsmodelle für beobachtbare Wartezeiten formulieren könnte.

Wie man sich leicht überlegen kann, ist die Charakterisierung von Wartezeitverteilungen durch einen Lage− und einen Streuungsparameter wenig sinnvoll: Einmal ist der Wertebereich von Wartezeiten beschränkt (nur positive Zahlen). Zum anderen ist eine Wartezeitverteilung häufig schief. Das gilt nicht für logarithmierte Wartezeiten: Hier reicht der Wertebereich von $-\infty$ bis $+\infty$ und die Verteilung ist häufig symmetrischer. Angenommen Wartezeiten sind exponentiell verteilt mit der Rate $r(t)=\lambda$. Die Dichtefunktion der Zufallsvariablen T lautet daher:

$$f(t) = \lambda \exp(-\lambda t)$$

Unter dieser Voraussetzung habe ich gezeigt (vgl. Abschnitt 2.3.3), daß der natürliche Logarithmus von T extremwertverteilt ist. Die Zufallsvariable $Y=\ln T$ hat daher die Dichtefunktion:

$$(5.6)\quad f(y) = \frac{1}{\sigma} \exp\left\{\frac{y-\mu}{\sigma} - \exp\frac{y-\mu}{\sigma}\right\} = \frac{1}{\sigma}\exp(w - \exp w)$$

$$\text{mit } w = \frac{y-\mu}{\sigma},\ \sigma = 1,\ \mu = -\ln\lambda$$

Die Verteilung hat die schöne Eigenschaft, zur Familie der "location−scale"−Verteilungen zu gehören (vgl. Abb. B.3). Man kann also die standardisierte, extremwertverteilte Zufallsvariable W dazu verwenden, um ein ähnliches Modell wie (5.1) für exponentiell verteilte Wartezeiten zu formu-

lieren. Der natürliche Logarithmus y dieser Wartezeiten t ist danach eine Funktion einer systematischen Komponente $\mu$ und einer Fehlerkomponente $\epsilon$:

$$y_i = \mu_i + \epsilon \qquad \epsilon = \sigma w \quad \text{und } W \sim E(0,1) \tag{5.7}$$

Der Fehlerterm besteht jetzt aus einer standardisierten extremwertverteilten Zufallsvariablen W mit konstanter Varianz $\sigma^2 = 1$.

Wenn man jetzt über vollständige Daten verfügt (d.h. jede Untersuchungseinheit hatte mindestens ein Ereignis), könnte man wie gehabt eine Regression der logarithmierten Zeitdauern auf die Kovariaten rechnen. Die LS-Schätzer der Regressionskoeffizienten sind unverzerrt, haben jedoch eine größere Varianz (d.h. sind weniger effizient), weil die Normalverteilungsannahme nicht mehr zutrifft. In der Regel ist aber ein Teil der Beobachtungen zensiert, so daß eine einfache Anwendung von LS-Schätzungen nicht möglich ist. Man verwendet daher auch bei diesem Modell in der Regel ML-Schätzungen.

Beschäftigen wir uns jedoch einmal mit den Annahmen dieses Modells: Zielvariable ist also die logarithmierte Wartezeit y, deren systematische Komponente $\mu$ auf Kovariaten zurückgeführt wird und deren Fehlerkomponente $\epsilon$ einer bestimmten Fehlerverteilung genügt (z.B. der Extremwertverteilung bei exponentiell verteilten Wartezeiten t):

$$y_i = \ln t_i = \mathbf{x}_i \boldsymbol{\alpha} + \epsilon \tag{5.8}$$

Durch erneute Exponentiation kann man die Logarithmierung wieder rückgängig machen:

$$t_i = \exp y_i = \exp(\mathbf{x}_i\boldsymbol{\alpha}) \exp\epsilon \tag{5.9}$$

Im Gegensatz zum Modell (5.2) proportionaler Risiken verwendet dieses Modell nicht die Rate, sondern die Wartezeiten als Zielvariable. Für einen Vergleich ist es jedoch interessant zu wissen, welche Ratengleichung das Modell (5.8) impliziert (vgl. dazu KALBFLEISCH/PRENTICE 1980: 33f. oder LAWLESS 1982: 276f.). Es gilt:

$$r(t^*|\mathbf{x}_i) = \lambda_0(t^*) \exp(-\mathbf{x}_i\boldsymbol{\alpha}) \qquad \text{mit } t^* = t \exp(-\mathbf{x}_i\boldsymbol{\alpha}) \tag{5.10}$$

Ein Vergleich dieser Funktionsgleichung mit (5.2) zeigt zunächst, daß sich die Effekte der Kovariaten $\exp(\mathbf{x}_i\beta)$ bzw. $\exp(-\mathbf{x}_i\alpha)$ bis auf das Vorzeichen nicht unterscheiden.[1] Weiterhin zeigt ein Vergleich der beiden Basisraten $\lambda_0(t)$ und $\lambda_0(t^*)$, daß die Kovariaten in Modell (5.8) wegen $t^* = t \exp(-\mathbf{x}_i\alpha)$ auch multiplikativ auf die Zeit t des Prozesses wirken. Mit anderen Worten, die Kovariaten bestimmen nicht nur die Rate, sondern verändern (skalieren) auch die Zeitachse für jede Untersuchungseinheit. Ich bezeichne daher dieses Regressionsmodell als ein *Skalierungsmodell* (engl. accelerated failure time model). Man kann zeigen, daß Modelle mit *exponentiell, Weibull- und Gompertz-verteilten Wartezeiten* sowohl zur Klasse der Modelle mit proportionalen Risiken als auch zur Klasse der Skalierungsmodelle gehören (KALBFLEISCH/PRENTICE 1980: 34f.).

### *5.1.4 Ein allgemeines Erklärungsmodell für Übergangsraten*

Ein Erklärungsmodell für Verlaufsdaten sollte so allgemein sein, daß man alle in der Einführung genannten Einflußfaktoren gemeinsam überprüfen kann. Im einzelnen sind das:

1. *Zustandsabhängigkeit* (Markovian state dependence): Die Wahrscheinlichkeit eines Ereignisses ist davon abhängig, welcher Zustand aktuell eingenommen wird. Das ist die Grundannahme eines Markov-Prozesses erster Ordnung.
2. Veränderungen mit der *Zustandsdauer* (duration dependence): Die Wahrscheinlichkeit eines Ereignisses hängt davon ab, seit wann der aktuelle Zustand eingenommen wird. Diese Zeitabhängigkeit ist ausschließlich eine Funktion des aktuellen Zustands. Deshalb spricht man von Semi-Markov-Prozessen.
3. *Vorgeschichte, Häufigkeit und Art früherer Ereignisse* (occurence dependence): Die Wahrscheinlichkeit eines Ereignisses ist nicht nur davon abhängig, welcher Zustand aktuell eingenommen wird, sondern auch davon, welche vorherigen Zustände (Vorgeschichte) aufgetreten sind. Die Berücksichtigung der Vorgeschichte führt zu Markov-Prozessen höherer Ordnung.

---

1) Inhaltlich kann man sich diese Feststellung folgendermaßen plausibel machen: Wenn eine Kovariate die Rate in (5.2) erhöht ($\beta > 0$), dann ereignen sich Veränderungen früher und die Wartezeit ist entsprechend kürzer. Die gleiche Kovariate muß daher in (5.8) einen negativen Effekt ($\alpha < 0$) auf die (logarithmierte) Wartezeit haben.

4. *Vorgeschichte, Dauer früherer Zustände* (lagged duration dependence): Die Wahrscheinlichkeit eines Ereignisses hängt von der Dauer früherer Zustände ab.
5. *Heterogenität*: Die Wahrscheinlichkeit eines Ereignisses hängt davon ab, welche Eigenschaften die Untersuchungseinheit aufweist, bei der das Ereignis auftritt.

Betrachten wir einmal unter diesen Gesichtspunkten das o.g. Modell (5.2) proportionaler Risiken:

$$r(t|\mathbf{x}_i) = \lambda_0(t) \exp(\mathbf{x}_i\beta)$$

$\lambda_0(t)$ ist eine zunächst ganz allgemein gehaltene Funktion zur Erfassung von Zeitabhängigkeiten. Wenn keine Kovariaten berücksichtigt werden, entspricht die Übergangsrate $\lambda_0(t)$. Daher der Name Basisrate. Sie mißt quasi das Risiko einer fiktiven Person, die bei allen Merkmalen den Wert Null hat. Je nachdem welche Restriktionen ihr auferlegt werden, ergeben sich spezifische zeitabhängige Modelle.

Wenn man annimmt, daß $\lambda_0(t)$ zwar unbekannt, aber für alle Untersuchungseinheiten gleich ist, dann ergibt sich das Regressionsmodell von COX (1972). Dieser Ansatz wird auch partiell parametrisch genannt, weil nur Teile des Modells spezifiziert werden. Er ist vor allen Dingen dann sinnvoll, wenn man keine Vorstellungen über die Art der Veränderung des untersuchten Prozesses hat. Er setzt allerdings spezifische Schätzverfahren voraus.

Wenn $\lambda_0(t)=\lambda$ einer Konstanten entspricht, ergibt sich das mittlerweile bekannte Modell eines Poisson-Prozesses mit exponentiell verteilten Wartezeiten. Nimmt man hingegen an, daß die Rate monoton im Zeitablauf sinkt (zunimmt), dann verwendet man für $\lambda_0(t)$ eine Funktion der Zeit t. Die *Weibull-* und die *Gompertzrate* sind Beispiele dieser Art. In diesem Fall gilt $\lambda_0(t)=\exp(\beta_0+\beta_1 \ln t)$ bzw. $\lambda_0(t)=\exp(\beta_0+\beta_1 t)$. Im Prinzip läßt sich aber jede beliebig komplizierte Funktion der Zeit verwenden, die aktuelle t und vergangene Zustandsdauern s enthält. Im Gegensatz zu dem Regressionsmodell von COX sind jetzt alle Parameter des Modells spezifiziert, so daß man in diesem Fall auch von *vollständig parametrischen Modellen* spricht.

Betrachten wir nun, welche Merkmale in dem Vektor $\mathbf{x}_i$ enthalten sein können. $\mathbf{x}_i$ umfaßt zunächst alle konstanten Charakteristika der Untersuchungseinheiten. Dadurch läßt sich die Heterogenität des Untersuchungs-

materials kontrollieren. Im Prinzip können sich diese Merkmale auch im Zeitablauf ändern (wie z.B. Einkommen[1]), wie ich jedoch später zeigen werde, ist bei Verwendung zeitabhängiger Kovariaten $\mathbf{x}_i(t)$ besondere Sorgfalt notwendig. Schließlich kann $\mathbf{x}_i$ auch Charakteristika der Vorgeschichte enthalten, so daß man z.B. überprüfen kann, ob die Rate mit der Häufigkeit früherer Ereignisse abnimmt.

Wenn man schließlich bedenkt, daß Zustandsabhängigkeiten dadurch erfaßt werden können, daß je ein Modell für jede Übergangsrate $r_{jk}(t)$ von j nach k formuliert wird, dann sind alle 5 genannten Einflußfaktoren abgedeckt.

(5.2c) $$r_{jk}(t|\mathbf{x}_{ijk}) = \lambda_{0jk}(t) \exp(\mathbf{x}_{ijk}\boldsymbol{\beta}_{jk})$$

Man beachte, daß Basisrate $\lambda_0(t)$, Kovariaten $\mathbf{x}_i$ und Parameter $\boldsymbol{\beta}$ jetzt je nach Übergang verschieden sein können. Das Regressionsmodell (5.2c) hat lediglich den Nachteil, keinen Fehlerterm $\epsilon(t)$ zu enthalten, um Meßfehler und unbekannte Einflüsse zu kontrollieren. Im Prinzip spricht jedoch nichts dagegen, einen solchen Term in (5.2) aufzunehmen. Man muß allerdings eine Annahme über die Verteilung und die Autokorrelation dieser Fehler $\epsilon(t)$ machen:

(5.2d) $$r_{jk}(t|\mathbf{x}_{ijk}) = \lambda_{0jk}(t) \exp(\mathbf{x}_{ijk}\boldsymbol{\beta}_{jk} + \epsilon_{jk}(t))$$

Die meisten praktisch verfügbaren Verfahren gehen jedoch davon aus, daß $\epsilon(t)$ Null ist.

Abschließend sei noch darauf hingewiesen, daß sich ein solches allgemeines Erklärungsmodell auch für zeitdiskrete Raten (2.9a) aufstellen läßt. Ich werde jedoch in der folgenden Auswertung kontinuierliche Daten voraussetzen. Einige Hinweise auf Regressionsmodelle für zeitdiskrete Daten finden sich in Kapitel 6.

---

1) An dem Beispiel Einkommen erkennt man, daß hiermit auch Merkmale gemeint sein können, die den Zustand (Tätigkeit) beschreiben, den die jeweilige Untersuchungseinheit aktuell einnimmt.

### *5.1.5 Überblick über dieses Kapitel*

In diesem Kapitel möchte ich mich hauptsächlich mit Regressionsmodellen für Raten unter der Annahme proportionaler Risiken beschäftigen, weil diese Modellklasse am häufigsten verwendet wird und der Benutzer auch hinreichende EDV-Unterstützung erhält. In den folgenden 6 Abschnitten möchte ich auf einige der oben aufgeworfenen Fragen ausführlicher eingehen.

Zunächst einmal erscheint es mir wichtig, das ML-Schätzprinzip an Hand eines sehr einfachen Beispiels zu demonstrieren. In Abschnitt 5.2 werde ich daher an Hand der MZ71-Daten untersuchen, inwieweit sich soziale Abstiege durch die dichotomisierte Variable Qualifikation prognostizieren lassen. Ich werde auch zeigen, in welche praktischen Schwierigkeiten man gerät, wenn man versucht, diese Fragestellung mit Hilfe einfacher LS-Schätzungen zu bearbeiten.

Wie man die Ergebnisse einer ML-Schätzung (möglichst anschaulich) interpretiert, möchte ich dann an einem etwas realistischeren Beispiel vorführen. In Abschnitt 5.3 werde ich den multivariaten Einfluß aller in Abschnitt 3.2 diskutierten Merkmale (außer der Zeitabhängigkeit) untersuchen. Dabei geht es um die Frage, ob die einzelnen Variablen die von mir erwarteten Effekte zeigen (Parametertests) und ob wirtschaftliche Randbedingungen (mehrere Variablen) berufliche Abstiege besser erklären als berufliche Aufstiege. Da mit der letzten Frage verschiedene Arten von Tätigkeitswechseln angesprochen sind, muß die vorher abgeleitete Likelihoodfunktion auf multiple Ereignisse erweitert werden.

Bis zu diesem Punkt gehe ich von einem zeitkonstanten Prozeß (d.h. exponentiell verteilten Wartezeiten) aus. Auf Grund der Überlegungen in Abschnitt 3.2 ist jedoch anzunehmen, daß sich verschiedene Arten von Tätigkeitswechseln auch in zeitlicher Hinsicht voneinander unterscheiden. In Abschnitt 5.4 werde ich daher die bisher an einigen Stellen nur kursorisch erwähnten zeitabhängigen Modelle ausführlich besprechen. Es zeigt sich, daß sich an dem bisherigen Schätzverfahren nichts wesentlich ändert, wenn die Form der Zeitabhängigkeit genau spezifiziert wird.

Von den zeitabhängigen Regressionsmodellen ist das Modell von COX (1972) besonders attraktiv, weil hier keine Aussage über die Form der Zeitabhängigkeit $\lambda_0(t)$ nötig ist. Damit ist das Modell jedoch nur teilweise spezifiziert und erfordert eine gewisse Abänderung des ML-Prinzips.

Abschnitt 5.5 beschäftigt sich mit diesen PL-Schätzungen (Partial Likelihood).

In Abschnitt 5.6 werden schließlich die verschiedenen Einzelergebnisse zusammengefaßt und die Auswahl eines passenden Modells für die untersuchten Daten unter allgemeinen Gesichtspunkten diskutiert.

Im ganzen Kapitel 5 gehe ich von einem Prozeß aus, in dem sich Ereignisse nicht wiederholen können. Alle empirischen Analysen beschränken sich daher ausschließlich auf die ersten Episoden der ausgewählten Personen. Wiederholbare Ereignisse und weiterführende Fragen der Verlaufsdatenanalyse (z.B. zeitabhängige Kovariate) sind Gegenstand von Kapitel 6.

## 5.2 *Maximum-Likelihood-Schätzung von Regressionsmodellen für Verlaufsdaten*

In diesem Abschnitt möchte ich an Hand eines ganz einfachen Beispiels mit einem singulären Ereignis und einer dichotomen Kovariaten das prinzipielle Vorgehen bei ML-Schätzungen demonstrieren. Abschnitt 5.2.1 enthält eine formale Definition des ML-Prinzips und eine Likelihood-Funktion für diesen einfachen Verlaufsprozeß. Likelihood-Funktionen für allgemeinere Prozesse werden in Abschnitt 5.3 diskutiert.

Normalerweise ist man bei ML-Schätzungen auf Computerprogramme angewiesen, die das Maximierungsproblem iterativ lösen. Bei diesem einfachen Beispiel ist es jedoch möglich, eine explizite Lösung anzugeben (vgl. Abschnitt 5.2.2). In diesem Fall kann man auch angeben, wie groß die Verzerrung der Ergebnisse wäre, wenn man zensierte Beobachtungen ignoriert oder als Quasi-Ereignisse behandelt (vgl. Abschnitt 5.2.3). Dieses Kapitel endet mit einer Diskussion der wesentlichen Teststatistiken (vgl. Abschnitt 5.2.4), die wiederum an Hand des Beispiels demonstriert werden (vgl. Abschnitt 5.2.5).

Für das Beispiel wurde die Variable "Ausbildung" der MZ71-Daten dichotomisiert und folgendermaßen kodiert: 0 = niedrige Qualifikation (Ausbildungsdauer bis 10 Jahre), 1 = hohe Qualifikation (Ausbildungsdauer 11 Jahre und mehr). Von den insgesamt 7587 Personen haben danach 4339 eine niedrige und 3248 eine hohe Qualifikation. Wir interessie-

ren uns nur für soziale Abstiege und behandeln alle anderen Tätigkeitswechsel als zensierte Beobachtungen (vgl. die explorativen Analysen in Kapitel 4).

### *5.2.1 Grundprinzip von ML−Schätzungen*

Ganz allgemein gesprochen, verwenden ML−Schätzungen die Werte als Schätzer für die unbekannten Parameter eines Regressionsmodells, bei denen sich am ehesten die vorliegende Stichprobe ergeben hätte. Ausgangspunkt ist ein Modell über die unbekannte Grundgesamtheit, an Hand dessen man in Form eines Planspiels überlegt, was unter den Annahmen des Modells passieren kann. Unter anderem kann man sich überlegen, wie die vorliegende Stichprobe zustandegekommen ist. Man kann sogar angeben, wie plausibel (likely, "wahrscheinlich") die Stichprobe unter den Modellannahmen ist. Wenn man jetzt die Parameter des Modells variiert (verschiedene Szenarien durchspielt), erkennt man, unter welchen Bedingungen die vorliegende Stichprobe ein plausibles Untersuchungsergebnis ist. Die Parameter, unter denen die Stichprobe am plausibelsten ist, sind wahrscheinlich die Bedingungen, unter denen die Stichprobe tatsächlich zustandegekommen ist. Man verwendet sie daher als Schätzer für die unbekannten Parameter der Grundgesamtheit. Dabei ist man implizit davon ausgegangen, daß das Modell richtig und die Daten repräsentativ sind.

Eine *ML−Schätzung* besteht also aus drei Schritten:

1. Formulierung eines Modells,
2. Berechnung der Likelihood der Stichprobe unter den Modellannahmen,
3. Maximierung dieser Likelihood.

Ein einfaches Modell für das Beispiel lautet: Das Abstiegsrisiko unterscheidet sich zwischen beiden Qualifikationsgruppen, aber Abstiege treten generell im Zeitablauf mit konstanter Rate auf (Exponentialverteilung). Die Basisrate $\lambda_0(t)$ ist also keine Funktion der Zeit, sondern entspricht einer Konstanten $\lambda_0(t)=\lambda_0=\exp\beta_0$. Die entsprechende Regressionsgleichung der Abstiegsrate der Person i lautet:

$$r(t|x_{i1}) = \exp(\beta_0 + \beta_1 x_{i1}) \tag{5.11}$$

oder in Vektorschreibweise:

$$r(t|\mathbf{x}_i) = \exp(\mathbf{x}_i\beta) \tag{5.12}$$

$\beta$ ist dabei ein Spaltenvektor mit den unbekannten Parametern $\beta_0$ und $\beta_1$. $\mathbf{x}_i$ ist ein Zeilenvektor der exogenen Variablen $x_{i0}$ und $x_{i1}$ für Person i. Dabei hat $x_{i0}$ für alle Beobachtungen den Wert 1 und $x_{i1}$ ist die Dummy-Variable "Qualifikation". Das Abstiegsrisiko für die beiden Qualifikationsgruppen lautet daher:

$$\begin{aligned} r_1(t) &= \lambda_1 = \exp(\beta_0 + \beta_1) \quad \text{(hohe Qualifikation: } x_{i1}=1) \\ r_0(t) &= \lambda_0 = \exp \beta_0 \quad \text{(niedrige Qualifikation: } x_{i1}=0) \end{aligned} \tag{5.13}$$

Die nächste Aufgabe besteht jetzt darin, die Likelihood der Stichprobe unter diesen Modellannahmen zu berechnen.

Angenommen alle $i=1,\ldots,N$ Personen hätten ein Ereignis innerhalb der Untersuchungsdauer. Betrachtet man die Wartezeit bis zum Eintritt eines solchen Ereignisses als diskrete Zufallsvariable, dann kann man direkt die Wahrscheinlichkeit $f(t_k)$ berechnen, mit der das Ereignis zum jeweiligen Zeitpunkt auftritt (vgl. Gleichung 2.1a). Geht man dagegen von einer kontinuierlichen Zufallsvariablen aus, dann betrachtet man alternativ die Dichte f(t) (vgl. Gleichung 2.1b).[1] Beide Funktionen lassen sich jedenfalls auf die Rate $q_k$ bzw. r(t) zurückführen (vgl. die Gleichungen 2.14 und 2.16). Wir wollen von einer kontinuierlichen Zufallsvariablen ausgehen und verwenden für r(t) das obige Regressionsmodell (5.11). Die Dichte wird damit eine Funktion der unbekannten Parameter $\beta$ und der Daten $\mathbf{x}_i$: $f(t_i|\beta,x_i)$. Bei Unabhängigkeit und identischer Verteilung der Beobachtungen ergibt sich die gesamte Likelihood L der Stichprobe aus dem Produkt der Einzeldichten:

$$L(\beta|t,\mathbf{X}) = \prod_{i}^{N} f(t_i|\beta,\mathbf{x}_i) \tag{5.14}$$

1) Der folgende Ausdruck (5.14) entspricht daher nur im diskreten Fall der Gesamtwahrscheinlichkeit der Stichprobe. Im allgemeinen Fall (unter Verwendung der Dichte) gebraucht man daher den Begriff "Likelihood", den ich an einigen Stellen umgangssprachlich mit "Plausibilität" übersetzt habe.

Diesen Ausdruck bezeichnet man auch als *Likelihood-Funktion*. Wenn man jetzt verschiedene Werte für $\beta_0$ und $\beta_1$ in (5.14) einsetzt, kann man die Plausibilität der Stichprobe unter verschiedenen Bedingungen durchspielen. Die einzelnen Beobachtungen $t_i$ und $\mathbf{x}_i$ werden als gegeben vorausgesetzt (genauso wie das eigentliche Modell) und die Likelihood ist nur noch eine Funktion der unbekannten Parameter (wie in der Klammer hinter L angedeutet).

Im dritten Schritt wird nun ein Set von Parametern $\hat{\beta}$ gesucht, der die Likelihood des Untersuchungsergebnisses maximiert, so daß gegenüber jedem anderen Set von Parametern $\beta$ gilt:

$$\ln L(\hat{\boldsymbol{\beta}}|\mathbf{t},\mathbf{X}) \geq \ln L(\bar{\boldsymbol{\beta}}|\mathbf{t},\mathbf{X}) \tag{5.15}$$

Dabei macht es keinen Unterschied, ob man die Likelihood-Funktion selbst oder ihren natürlichen Logarithmus betrachtet. Im letzteren Fall werden die folgenden mathematischen Operationen lediglich etwas einfacher (Produkte werden zu Summen). Natürlich kann man versuchen, durch geschicktes Probieren diese Werte zu finden. Eine mathematisch elegante Lösung ergibt sich jedoch durch *partielle Differentiation* der logarithmierten Likelihoodfunktion:[1]

$$\Theta_p(\hat{\beta}) = \frac{\partial \ln L(\boldsymbol{\beta}|\mathbf{t},\mathbf{X})}{\partial \beta_p} = 0 \qquad \text{für } p = 0,1 \tag{5.16}$$

Die ersten partiellen Ableitungen $\Theta_p(\hat{\beta})$ werden auch häufig als *Scores* bezeichnet. Man kann sie der Reihe nach in einen Vektor $\Theta(\hat{\beta})$ schreiben, den man den *Gradienten* der Likelihood-Funktion nennt. Durch Nullsetzen der Scores erhält man die gesuchten ML-Schätzer.

---

1) Man erinnere sich dazu noch einmal an die Schulmathematik: Dort ging es darum, die Nullstellen einer Funktion zu bestimmen. Um sicherzustellen, daß es sich dabei um keine Wendepunkte handelt, mußte zusätzlich geprüft werden, ob die 2. (partiellen) Ableitungen ungleich 0 sind. Im allgemeinen Fall führt das zu der Forderung, daß die Matrix der 2. (partiellen) Ableitungen, die *Hesse'sche Matrix*, negativ definit sein muß. Falls ML-Programme eine entsprechende (schwer verständliche) Fehlermeldung ausgeben, dann ist das ein Hinweis darauf, daß sie kein globales Maximum finden können. In einigen Fällen ist es sogar unmöglich, die 1. und 2. Ableitungen einer Likelihood-Funktion anzugeben. In diesem Fall ist man auf andere Suchstrategien angewiesen. Weitere Hinweise findet man dazu in einem Fachbuch zur numerischen Optimierung (z.B. CHAMBERS 1977, MURRAY 1972 oder WALSH 1975).

Soweit das Grundprinzip – jedoch enthalten Verlaufsdaten in der Regel auch zensierte Beobachtungen. Um diesem Umstand gerecht zu werden, muß die Likelihood–Funktion (5.14) entsprechend erweitert werden. Jede Stichprobe besteht im Prinzip aus zwei Teilen: Die Menge $D$ der Beobachtungen mit Ereignis, sowie die Menge $C$ der Beobachtungen, von denen man nur weiß, daß sie bis zum Zeitpunkt der Zensierung ohne Ereignis überlebt haben.

Nach unseren Überlegungen über rechtszensierte Beobachtungen (vgl. Abschnitt 2.5.2.2) kann die Likelihood dieser "zweigeteilten" Daten unter Verwendung einer getrennten (Status–)Variablen $\delta$ wie folgt beschrieben werden:

$$L(\boldsymbol{\beta}|\boldsymbol{\delta},\mathbf{t},\mathbf{X}) = \prod_{i}^{N} f(t_i|\boldsymbol{\beta},\mathbf{x}_i)^{\delta_i}\, S(t_i|\boldsymbol{\beta},\mathbf{x}_i)^{1-\delta_i} \tag{5.17}$$

$$\text{mit } \delta_i = \begin{cases} 1 \text{ (Ereignis)} \\ 0 \text{ (Zensierung)} \end{cases}$$

und unter Verwendung von (2.14) gilt weiter:

$$L(\boldsymbol{\beta}|\boldsymbol{\delta},\mathbf{t},\mathbf{X}) = \prod_{i}^{N} S(t_i|\boldsymbol{\beta},\mathbf{x}_i)\, r(t_i|\boldsymbol{\beta},\mathbf{x}_i)^{\delta_i} \tag{5.18}$$

Dies ist die *allgemeine Likelihood–Funktion für Prozesse mit singulären Ereignissen.* Im Gegensatz zu (5.14) bestehen die Stichprobeninformationen jetzt aus $\mathbf{t}$, $\mathbf{X}$ und der Statusvariablen $\boldsymbol{\delta}$. Durch Einsetzen eines beliebigen (vollständig parametrischen) Regressionsmodells für die Rate r(t) kann man die unterschiedlichsten Prozesse modellieren und dabei zensierte und unzensierte Beobachtungen gleichermaßen berücksichtigen.

Bevor ich das Ergebnis der ML–Schätzung für das obige Beispiel demonstriere, sei noch darauf hingewiesen, daß es sich bei dieser Likelihood–Funktion strenggenommen um die Likelihood für *Typ I zensierte Prozesse* handelt. Für *Typ II zensierte Prozesse* ergibt sich zwar im Prinzip die gleiche Funktion, jedoch sind die Mengen $D$ und $C$ von vorneherein festgelegt.

Man braucht daher die Statusvariable $\delta$ nicht und kann gleich die entsprechenden zu Beginn der Untersuchung festgelegten Zahlen einsetzen. Angenommen man beendet die Untersuchung, nachdem D von N Beobachtungen ein Ereignis hatten. Die verbleibenden $C = N - D$ Beobachtungen werden dann zum Zeitpunkt des D−ten Ereignisses zensiert − oder anders ausgedrückt, haben eine Überlebenszeit von $t_D$ Zeiteinheiten. Wenn man noch berücksichtigt, daß die D Ereignisse in unterschiedlicher Reihenfolge möglich sind, dann lautet die allgemeine Likelihood−Funktion entsprechend:

$$(5.19) \qquad L(\boldsymbol{\beta}|\mathbf{t},\mathbf{X}) = \frac{N!}{C!} \prod_{i \in C} S(t_D|\boldsymbol{\beta},\mathbf{x}_i) \prod_{i \in D} f(t_i|\boldsymbol{\beta},\mathbf{x}_i)$$

Wenn man Gleichung (5.17) und (5.19) vergleicht, sieht man, daß sie sich im Prinzip nur durch den Faktor N!/C! unterscheiden. Dieser Faktor hat jedoch auf das Maximierungsproblem keinen Einfluß. Etwas anders formuliert kann man sagen, daß bei Typ II zensierten Prozessen nur eine Zufallsvariable (nämlich die Wartezeit) beobachtet wird, während es bei Typ I zensierten Prozessen zwei Zufallsvariablen sind (die Wartezeit und die Statusvariable). An dieser Stelle wird plausibel, warum Schätz− und Inferenzprobleme bei Typ II zensierten Prozessen sehr viel einfacher zu handhaben sind und warum man bei Typ I zensierten Prozessen die *Annahme unabhängiger Zensierung* machen muß − insbesondere dann, wenn die Beobachtungsdauer individuell variiert und nicht für alle Untersuchungseinheiten einheitlich festliegt (z.B. bei *progressiver Zensierung*). Bei Typ II zensierten Prozessen ist es möglich, für beliebige Modelle exakte Konfidenzintervalle der gefundenen Schätzer anzugeben, während man sich bei Typ I zensierten Prozessen häufig auf allgemeine Resultate der ML−Theorie beziehen muß (LAWLESS 1982: 43f.).

### *5.2.2 Ein Beispiel mit einer dichotomen Kovariaten*

In diesem Abschnitt möchte ich zeigen, wie man mit Hilfe der Likelihood−Funktion (5.18) Schätzwerte für die beiden Parameter des Regressionsmodells (5.11) findet. Um die Ableitungen nicht unnötig zu komplizieren, verwende ich an einigen Stellen eine Umformung des Modells:

$$(5.20) \qquad r(t|x_{i1}) = \lambda_i = \exp(\beta_0 + \beta_1 x_{i1}) = \exp \beta_0 \exp(\beta_1 x_{i1}) = a_0 a_1^{x_{i1}}$$

$a_0 = \exp\beta_0$ und $a_1 = \exp\beta_1$ sind dabei die Antilogarithmen der unbekannten Parameter. Wie sich später zeigt (vgl. Abschnitt 5.3), eignen sie sich besonders zur Interpretation log-linearer Abhängigkeiten.

Da das Modell von exponentiell verteilten Wartezeiten ausgeht, kann man die entsprechenden Formeln für r(t) und S(t) (vgl. Gleichung 2.17) in die allgemeine Likelihood-Funktion (5.18b) einsetzen:

$$L = \prod_{i}^{N} [\lambda_i^{\delta_i} \exp(-\lambda_i t_i)] \tag{5.21}$$

In dem Beispiel haben $N_0 = 4339$ Personen ein niedriges und $N_1 = 3248$ Personen ein hohes Qualifikationsniveau. Dementsprechend kann man die Likelihood-Funktion in zwei Teile aufsplitten und die entsprechenden Raten (5.20) der beiden Gruppen für $\lambda_i$ einsetzen.

$$L = \prod_{i}^{N_0} [(\exp \beta_0)^{\delta_i} \exp(-t_i \exp \beta_0)] \cdot \prod_{i}^{N_1} [(\exp \beta_0 \exp \beta_1)^{\delta_i} \exp(-t_i \exp \beta_0 \exp \beta_1)] \tag{5.22}$$

ML-Schätzer für $\beta_0$ und $\beta_1$ findet man nun durch Maximierung der logarithmierten Likelihood-Funktion:

$$\ln L = \sum_{i}^{N_0} (\delta_i \beta_0 - t_i \exp \beta_0) + \sum_{i}^{N_1} [\delta_i(\beta_0 + \beta_1) - t_i \exp \beta_0 \exp \beta_1] \tag{5.23}$$

Das Maximierungsproblem löst man durch partielle Differentiation von (5.23):

$$\frac{\partial \ln L}{\partial \beta_0} = \sum_{i}^{N_0} (\delta_i - t_i \exp \beta_0) + \sum_{i}^{N_1} (\delta_i - t_i \exp \beta_0 \exp \beta_1) \tag{5.24a}$$

$$= \sum_{i}^{N_0} \delta_i + \sum_{i}^{N_1} \delta_i - \exp \beta_0 \sum_{i}^{N_0} t_i - \exp \beta_0 \exp \beta_1 \sum_{i}^{N_1} t_i$$

(5.24b)
$$\frac{\partial \ln L}{\partial \beta_1} = \sum_i^{N_1} (\delta_i - t_i \exp \beta_0 \exp \beta_1)$$
$$= \sum_i^{N_1} \delta_i - \exp \beta_0 \exp \beta_1 \sum_i^{N_1} t_i$$

An der Stelle eines globalen Maximums sind diese 1. (partiellen) Ableitungen jeweils Null. Nach Nullsetzen und einigen Umformungen ergeben sich die ML−Schätzer für $\beta_0$ und $\beta_1$. Das Ergebnis läßt sich jedoch besser in Form der Antilogarithmen $a_0$ und $a_1$ festhalten:

(5.25a)
$$\exp \hat{\beta}_0 = \hat{a}_0 = \frac{\sum_i^{N_0} \delta_i}{\sum_i^{N_0} t_i} = \frac{D_0}{\Upsilon_0}$$

(5.25b)
$$\exp \hat{\beta}_1 = \hat{a}_1 = \frac{\sum_i^{N_1} \delta_i \sum_i^{N_0} t_i}{\sum_i^{N_0} \delta_i \sum_i^{N_1} t_i} = \frac{D_1 \Upsilon_0}{D_0 \Upsilon_1}$$

Durch Einsetzen in das Regressionsmodell (5.20) erhält man die Abstiegsraten der beiden Subgruppen:

(5.26a)
$$\hat{\lambda}_0 = \exp \hat{\beta}_0 = \hat{a}_0 = \frac{D_0}{\Upsilon_0}$$

(5.26b)
$$\hat{\lambda}_1 = \exp(\hat{\beta}_0 + \hat{\beta}_1) = \hat{a}_0 \hat{a}_1 = \frac{D_1}{\Upsilon_1}$$

Sie entsprechen genau dem Verhältnis aller Ereignisse $D_j$ in der jeweiligen Subgruppe j bezogen auf die Gesamtdauer $\Upsilon_j$ aller (zensierter und unzensierter) Wartezeiten in der gleichen Subgruppe.

In diesem einfachen Beispiel genügt also eine Auszählung der Ereignisse und der Gesamtdauer der Wartezeiten in beiden Subgruppen, um die Parameter des Modells bzw. die Abstiegsrisiken zu berechnen. Nach den

Daten in Tabelle 5.1 beträgt die Abstiegsrate bei niedriger Qualifikation 0,014 Abstiege pro Jahr Berufstätigkeit. Dieser Wert ist etwas höher als bei Personen mit hoher Qualifikation. Dort ereignen sich nur 0,009 Abstiege pro Jahr Berufstätigkeit. Das Regressionsmodell lautet dementsprechend:

$$(5.27) \qquad i(t|x_{i1}) = \exp(-4{,}286 - 0{,}463x_{i1}) = 0{,}014 \cdot 0{,}630^{x_{i1}}$$

*Tabelle 5.1: Häufigkeit und Dauer von Tätigkeiten nach Art des Tätigkeitswechsels und Qualifikation*

| Tätigkeit | | Alle Personen | | Personen mit niedriger Qualifikation | | Personen mit hoher Qualifikation | |
|---|---|---|---|---|---|---|---|
| erste | N | 7587 | 100 % | 4339 | 100 % | 3248 | 100 % |
| Tätigkeiten | ϒ | 58969 | 100 % | 35184 | 100 % | 23785 | 100 % |
| Tätigkeiten | D | 690 | 9.1 % | 484 | 11.2 % | 206 | 6.3 % |
| mit Abstieg | U | 3453 | 5.9 % | 2491 | 7.1 % | 962 | 4.0 % |
| restliche | C | 6897 | 90.9 % | 3855 | 88.8 % | 3042 | 93.7 % |
| Tätigkeiten | V | 55516 | 94.1 % | 32693 | 92.9 % | 22823 | 96.0 % |

Nur erste Tätigkeiten. N, C, D Häufigkeit. ϒ, V, U Gesamtdauer in Jahren

Der Parameter $a_0$ entspricht der Abstiegsrate der Personen mit niedriger Qualifikation und $\beta_0$ ist der natürliche Logarithmus dieser Rate. Der Parameter $\beta_1$ mißt die Differenz der beiden logarithmierten Abstiegsraten, während sein Antilogarithmus $a_1$ das Verhältnis beider Raten erfaßt. Anders ausgedrückt, $a_1$ ist der Faktor, mit der die Abstiegsrate der Gruppe mit niedriger Qualifikation multipliziert werden muß, um die Abstiegsrate der Gruppe mit hoher Qualifikation zu erhalten.

Dieses Beispiel mit zwei Gruppen kann auf eine beliebige Anzahl Gruppen (d.h. eine polytome Kovariate) erweitert werden und es ergeben sich ähnlich einfache Schätzer. Bei der Analyse multivariater Zusammenhänge, insbesondere mit metrischen Kovariaten, ist es jedoch in der Regel unmöglich, eine explizite Lösung des Maximierungsproblems anzugeben. In diesem Fall benötigt man iterative Computerprogramme, die mit Hilfe von Verfahren numerischer Optimierung Maxima der Likelihood-Funktion finden.

### *5.2.3 Berücksichtigung zensierter Beobachtungen*

Betrachten wir nun einen Forscher, dem die zuvor beschriebenen Schätzmethoden unbekannt sind. Wenn das Abstiegsrisiko mit der Qualifikation variieren soll, dann ist die Wartezeit bis zu einem Abstieg um so kürzer, je weniger die Untersuchungsperson qualifiziert ist. Aus diesem Grund denkt unser Forscher an eine Regression der Wartezeiten auf die Dummy–Variable "Qualifikation". Bei der Verwendung der logarithmierten Wartezeiten entspricht dieses Vorgehen dem Skalierungsmodell (5.8). Unser Forscher steht jedoch vor einem praktischen Problem: Wie soll er die Wartezeiten behandeln, die in einem Untersuchungszeitraum nicht mit einem Abstieg enden (zensierte Beobachtungen)? Er erwägt zwei Auswertungsstrategien: a) Ignorierung der zensierten Beobachtungen durch ausschließliche Verwendung der Abstiege, b) Berücksichtigung der zensierten Beobachtungen als "Ereignisse" durch die Annahme eines Abstiegs am Ende des Untersuchungszeitraums. Es ist anzunehmen, daß beide Strategien die Ergebnisse verfälschen.

Angenommen man verwendet lediglich die unzensierten Beobachtungen (Strategie a). Für diesen reduzierten Datensatz ist folgender Ausdruck ein ML–Schätzer der beiden subgruppenspezifischen Abstiegsraten:

(5.28a) $$\tilde{\lambda}_0 = \frac{D_0}{U_0} = 0{,}194$$

(5.28b) $$\tilde{\lambda}_1 = \frac{D_1}{U_1} = 0{,}214$$

Ereignisse werden jetzt nur noch auf die Gesamtdauer $U_j$ aller unzensierten Wartezeiten bezogen. Die Antilogarithmen des dazugehörenden Regressionsmodells lauten dementsprechend:

(5.29a) $$\tilde{a}_0 = \frac{D_0}{U_0} = 0{,}194 \qquad \text{mit } \tilde{\beta}_0 = \ln \tilde{a}_0 = -1.638$$

(5.29b) $$\tilde{a}_1 = \frac{D_1\, U_0}{Y_0\, U_1} = 1{,}102 \qquad \text{mit } \tilde{\beta}_1 = \ln \tilde{a}_1 = 0.097$$

Die Verhältnisse haben sich jetzt total umgekehrt: Das Abstiegsrisiko der Personen mit hoher Qualifikation ist um einen Faktor 1,102 höher als das Abstiegsrisiko der Personen mit niedriger Qualifikation (Vergleichsgruppe).

Da hier ein Teil der Stichprobe nicht berücksichtigt wird, müssen die Schätzer der Regressionsparameter verfälscht sein (*Auswahlfehler*). In diesem einfachen Beispiel kann man sich sogar genau überlegen, wie groß das Ausmaß der Verzerrung ist. Der Antilogarithmus $\tilde{a}_0$ in diesem reduzierten Datensatz ergibt sich, indem man den entsprechenden Antilogarithmus des vollständigen Datensatzes durch den Anteil unzensierter Wartezeiten in der Vergleichsgruppe dividiert:

$$\tilde{a}_0 = \hat{a}_0 \,/\, \frac{U_0}{\Upsilon_0} = \hat{a}_0 \,/\, 0{,}071 \tag{5.30}$$

Mit anderen Worten, je größer das Ausmaß der Zensierung in der Vergleichsgruppe, um so mehr überschätzt der Antilogarithmus $\tilde{a}_0$ den wahren Wert. Da in dem Beispiel nur 7,1% aller Wartezeiten in der Vergleichsgruppe mit einem Ereignis enden, ist der Antilogarithmus also ca. 14-mal größer als der tatsächliche Wert.

Bei dem Antilogarithmus $\tilde{a}_1$ sind die Verhältnisse etwas komplizierter:

$$\tilde{a}_1 = \hat{a}_1 \frac{\frac{U_0}{\Upsilon_0}}{\frac{U_1}{\Upsilon_1}} = \hat{a}_1 \frac{0{,}071}{0{,}040} \tag{5.31}$$

Hier hängt die Verzerrung davon ab, wie stark die Beobachtungen in beiden Gruppen zensiert sind. Ist beispielsweise das Ausmaß der Zensierung bei den höher Qualifizierten größer als in der Vergleichsgruppe, dann überschätzt der entsprechende Antilogarithmus ebenfalls den wahren Wert. Genauso ist es in dem Beispiel: Nur 4% aller Wartezeiten enden bei den höher Qualifizierten mit einem Ereignis, während es in der Vergleichsgruppe zumindest 7,1% sind. Dementsprechend wird der Antilogarithmus um einen Faktor 1,78 überschätzt.

Man kann sich die Richtung der Verzerrung auch inhaltlich überlegen:[1] Wenn es so ist, daß vermehrte Ausbildung das Risiko eines Abstiegs verringert, dann werden Abstiege viel eher bei Personen mit niedriger Qualifikation auftreten. Innerhalb eines begrenzten Untersuchungszeitraums ist somit die Wahrscheinlichkeit viel größer, bei Personen mit niedriger Qualifikation einen Abstieg zu beobachten. Mit anderen Worten, weniger Beobachtungen werden in der Vergleichsgruppe zensiert sein.

Zusammengefaßt kann man daher sagen, daß immer dann, wenn eine Variable einen Effekt auf die untersuchten Ereignisse hat, das Ausmaß zensierter Beobachtungen (notwendigerweise) systematisch mit den jeweiligen Ausprägungen der unabhängigen Variablen variiert. Ein positiver Effekt auf die Rate ($a > 1$) bedeutet geringere Zensierung bei hohen Variablenwerten, so daß Schätzungen, die ausschließlich unzensierte Beobachtungen verwenden, kleiner sind als die wahren Parameter. Für negative Variableneffekte ($a < 1$) kehrt sich diese Beziehung genau um. So auch in dem Beispiel: Hier hat das Merkmal Qualifikation einen negativen Effekt auf die Rate ($a_1 = 0{,}630$). Daher ist der Schätzer, der sich bei ausschließlicher Berücksichtigung unzensierter Beobachtungen ergibt, größer als der wahre Wert. In diesem Fall ist die Verzerrung sogar so groß, daß sich die Richtung des Variableneffektes umkehrt[2].

Eine Regression der 690 (logarithmierten) Wartezeiten auf die Dummy-Variable "Qualifikation" $x_{i1}$ ergibt übrigens folgende Schätzung:

$$\ln \bar{t}_i = \bar{\alpha}_0 + \bar{\alpha}_1 x_{i1} = 1{,}44 - 0{,}14 x_{i1}$$

Bei dem Vergleich mit den zuvor berechneten ML-Schätzern (5.29) sind zwei Dinge zu berücksichtigen: Erstens handelt es sich um Koeffizienten eines Skalierungsmodells und zweitens werden sie durch eine OLS-Schätzung bestimmt. Wie in 5.1.3 besprochen, unterscheiden sich die Parameter $\alpha$ des Skalierungsmodells nur durch das Vorzeichen von den Parametern $\beta$ des Ratenmodells. Es wären also 1,638 und 1,44 (Regressionskonstanten) bzw. 0,097 und 0,14 (Effekt Qualifikation) miteinander zu vergleichen.

---

1) Diese Überlegung ist für den multivariaten Fall wichtig, wo man das Ausmaß der Verzerrung nicht exakt berechnen kann, aber zumindest die Richtung angeben möchte.

2) Ähnliche Überlegungen kann man anstellen, wenn zensierte Beobachtungen nicht ignoriert, sondern als Quasi-Ereignisse (Ereignis zum Zeitpunkt der Zensierung) behandelt werden. Als kleine Übung kann man sich ja mal überlegen, welche Verzerrungen bei dieser Strategie entstehen.

Daß sie numerisch nicht exakt übereinstimmen, hängt damit zusammen, daß die $\tilde{\alpha}$ mit LS geschätzt wurden und sich daher gewisse Effizienzverluste ergeben. Die Abweichungen sind jedoch geringfügig. Wichtig ist lediglich, daß auch hier der Effekt der Qualifikation in die erwartete, falsche Richtung zeigt: Bei höherer Qualifikation ist die (logarithmierte) Zeitdauer bis zu einem Abstieg kürzer. Anders ausgedrückt, die LS−Schätzung ausschließlich unzensierter Wartezeiten vermittelt den irreführenden Eindruck, daß sich bei höherer Qualifikation schneller berufliche Abstiege ereignen.

Die mangelnde Berücksichtigung von zensierten Beobachtungen ist natürlich ein klarer Nachteil von LS−Schätzungen, so daß man in der Regel ML−Schätzungen verwenden wird. Die einzige Ausnahme ist, wenn für jede Kombination der Kovariaten mehrere (zensierte und unzensierte) Wartezeiten vorliegen. Für eine solche *Gruppe* von Beobachtungen kann man in einem ersten Schritt irgendeine Verlaufsstatistik berechnen, die zensierte und unzensierte Beobachtungen gleichermaßen berücksichtigt (z.B. Rate oder Überlebenswahrscheinlichkeit). In einem zweiten Schritt wird diese Statistik dann als Zielvariable in einer ganz gewöhnlichen LS−Regression verwendet, wobei man zur Verbesserung der Schätzer eine gewichtete Regression verwenden sollte. Dieses Vorgehen entspricht der Analyse tabellierter Daten (vgl. auch Abschnitt 6.4): Dabei unterteilen die Kovariaten die Untersuchungsgruppe in mehrere homogene Subpopulationen, für die immer mehrere Beobachtungen der Zielvariablen vorliegen (vgl. Tabelle 3.8). Bei metrischen Kovariaten (mit vielen Ausprägungen) ist das nicht immer der Fall, einige Anwendungsbeispiele findet man jedoch bei LAWLESS (1982: 295ff., 306ff. und 328ff.). Einen anderen Ansatz wählt MILLER (1976): Er verwendet in einer LS−Schätzung zensierte und unzensierte Beobachtungen, korrigiert aber die zensierten Beobachtungen durch eine besondere Gewichtungsprozedur. Einen Überblick über diesen und verwandte Ansätze liefert PÖTTER (1989).

### *5.2.4 Signifikanztests bei ML−Schätzungen*

Abschließend stellt sich die Frage, ob sich die Abstiegsrisiken der beiden Qualifikationsgruppen signifikant unterscheiden. Dazu benötigt man Angaben über den Stichprobenfehler der ML−Schätzer. Die ML−Theorie bietet dazu verschiedene Testverfahren an, die für gewisse Standardsituationen entwickelt wurden (z.B. das lineare Modell) und alle nur asymptotisch gelten. Zu diesen Standardsituationen gehören die hier betrachteten

(nicht-linearen) Modelle mit zensierten Beobachtungen zunächst einmal nicht. Die Optimalität der Testverfahren ist daher noch in einigen Fällen nachzuweisen. Asymptotisch besagt, daß die Verteilung der Teststatistiken nur in großen Stichproben gesichert ist, während sie bei kleinen Stichproben ungewiß ist. Dabei stellt sich natürlich die praktische Frage, was sind große und kleine Stichproben?

Diese Frage ist schwierig zu beantworten, weil das Problem nicht in allen Aspekten erforscht ist. Einige Simulationsexperimente mit den hier diskutierten Modellen für Verlaufsdaten zeigen jedoch, daß die berechneten Statistiken auch in Stichproben geringen Umfangs (N=25 oder 50) nicht zu abweichende Ergebnisse liefern (vgl. CARROLL et al. 1978, FENNELL et al. 1977, TUMA/HANNAN 1978). Solange keine gegenteiligen Studien vorliegen, gehe ich daher davon aus, daß die Testannahmen in einer breiten Palette von Anwendungsfällen zutreffen. Allerdings sollte man nicht mit zu *kleinen Stichproben* arbeiten, die weniger als 30 Beobachtungen umfassen. Außerdem sollte eine ausreichende Anzahl von vollständigen Beobachtungen (mind. 25 Ereignisse) vorliegen. Beschäftigen wir uns daher mit den Testverfahren im einzelnen.

Im Prinzip stehen drei verschiedene Testkriterien zur Verfügung:

a) Tests mit den Standardfehlern der ML-Schätzer,

b) Tests mit den Scores (Lagrange Multiplikator Tests)

c) Likelihood-Verhältnis-Tests.

Ich werde die drei Tests in dieser Reihenfolge besprechen und dann ein paar allgemeine Bemerkungen über ihre relativen Vor- und Nachteile machen, ehe ich ihre Anwendung an Hand des Beispiels demonstriere.

Zunächst kann man unter einigermaßen allgemeinen Bedingungen zeigen, daß die ML-Schätzer (5.16) normalverteilt sind mit Erwartungswert $\beta$ und Varianz-Kovarianz-Matrix $\mathbf{COV}(\beta)$, auf die ich später noch zu sprechen komme (vgl. Abschnitt 5.2.5):[1]

$$\hat{\beta} \simeq N(\beta, \mathbf{COV}(\beta)) \tag{5.32}$$

Anders ausgedrückt, die ML-Schätzer stimmen im Mittel mit den wahren Werten überein und der Stichprobenfehler ist berechenbar. Solche Über-

1) Alle folgenden Ableitungen gelten natürlich in gleicher Weise für die ML-Schätzer $\hat{\alpha}$ eines Skalierungsmodells.

legungen sind aus klassischen Regressionsanalysen bekannt, von daher wird diese Teststatistik am häufigsten verwendet. Mit ihr läßt sich prüfen, ob einzelne Parameter signifikant von Null verschieden sind. Mit Hilfe der Methode der linearen Kontraste läßt sich dieses Verfahren verallgemeinern, um komplexere Hypothesen zu testen, die mehrere Parameter betreffen. Die resultierende Teststatistik ist eine sogenannte *Wald–Statistik* und der allseits bekannte Signifikanztest eines Regressionskoeffizienten ist quasi ein Spezialfall.

Anstatt der ML–Schätzer kann man auch jeden anderen Parameterwert $\tilde{\beta}$ (oder eine Menge von Parameterwerten $\tilde{\beta}$) in die Score–Funktionen einsetzen, wenn man annimmt, daß er (sie) die Verhältnisse in der Grundgesamtheit richtig charakterisiert(en). Unter dieser Nullhypothese kann man die Scores bzw. die Varianz–Kovarianz–Matrix erneut berechnen und es läßt sich zeigen, daß die folgende sogenannte *Score–Statistik* näherungsweise Chi–Quadrat–verteilt ist:

$$(5.33a) \qquad S^2 = \Theta'(\tilde{\beta})\mathbf{COV}(\tilde{\beta})\Theta(\tilde{\beta}) \qquad \text{mit } S^2 \simeq \chi^2_q$$

Dabei ist der Spaltenvektor $\Theta(\tilde{\beta})$ aller Score–Funktionen. Die Anzahl der Freiheitsgrade df entspricht der Anzahl q der getesteten Parameter. Dieses Verfahren ist der Methode der linearen Kontraste sehr ähnlich, außer daß hier die Varianz–Kovarianz–Matrix für die in der Nullhypothese angenommenen Parameter neu berechnet wird. Auch hier kann man sowohl einzelne Parameter als auch Parametergruppen testen.

Für unsere Zwecke genügt folgender Spezialfall des Score–Testes: Beschreiben die Parameter $\tilde{\beta}$ der Nullhypothese die Grundgesamtheit angemessen, dann sollten die Scores für diese Parameter Null sein. Dementsprechend kann man zeigen, daß die Scores näherungsweise normalverteilt sind mit Erwartungswert 0:

$$(5.33b) \qquad \Theta(\tilde{\beta}) \simeq N(0,\mathbf{COV}(\beta)^{-1})$$

Die Varianz–Kovarianz–Matrix der Scores entspricht genau der inversen Matrix $\mathbf{COV}(\beta)^{-1}$ der ML–Schätzer.

Schließlich kann man noch die globalen Werte der Likelihood–Funktionen verschiedener Modelle vergleichen, die sich nur durch die Parameter unterscheiden, deren Signifikanz man überprüfen möchte. Je nachdem, ob sich die Modelle durch einen oder mehrere Parameter unterscheiden, erge-

ben sich Signifikanztests für einzelne Parameter oder Parametergruppen. Voraussetzung für diesen sogenannten *Likelihood-Verhältnis-Test* ist allerdings, daß es sich um hierarchische Modelle handelt. D.h. das Modell a, das nur p Parameter verwendet, muß ein Submodell des umfassenderen Modells A mit $s=p+q$ Parametern sein. Unter der Voraussetzung, daß die q zusätzlich geschätzten Parameter Null sind (Nullhypothese), ist das folgende Verhältnis der beiden Likelihood-Werte näherungsweise Chi-Quadrat-verteilt:

$$X^2 = 2 \ln \frac{L(A)}{L(a)} \qquad \text{mit } X^2 \simeq \chi^2_q \tag{5.34}$$

Implizit sind die q Parameter natürlich in Modell a enthalten, sie sind dort nur alle gleich Null. Das legt folgende Verallgemeinerung nahe: Unter der Annahme, daß q Parameter eines Modells a bestimmte feste Werte haben (nicht notwendigerweise Null), ist das Verhältnis der Likelihood dieses Modells zur Likelihood des gleichen Modells A, in dem diese q Parameter frei variieren dürfen, Chi-Quadrat-verteilt.

Ehe man diese vielen Teststatistiken anwendet, möchte man natürlich wissen, was ihre relativen Vorteile sind. Zunächst ist festzustellen, daß alle drei Statistiken nicht notwendigerweise gleiche Ergebnisse liefern. Das ist nur dann der Fall, wenn die Likelihood-Funktion gewissen Normalitätsbedingungen genügt. Insbesondere kann man zeigen, daß die Wald-Statistik gegenüber Reparametrisierungen (z.B. standardisierte statt unstandardisierte Regressionskoeffizienten) nicht invariant ist. Darüber hinaus gelten die Verteilungsannahmen, wie oben schon erwähnt, nur in Standardsituationen sowie in großen Stichproben. Für Standardmodelle hat sich jedoch gezeigt, daß die Likelihood-Verhältnis-Statistik auch in kleinen Stichproben noch am ehesten diesen Annahmen gerecht wird. Von daher gibt es gewisse Argumente für eine Bevorzugung der Statistik (5.34). Am einfachsten lassen sich Konfidenzintervalle nach (5.32) berechnen, dagegen benötigt man für (5.33) und (5.34) häufig mehr Rechenzeit. Die Wald-Statistik ist deshalb einfacher zu berechnen, weil sich die Varianz-Kovarianz-Matrix quasi als Nebenprodukt des Maximierungsproblems ergibt (s. folgender Abschnitt). Dagegen muß sowohl für die Likelihood-Verhältnis- als auch für die Score-Statistik immer ein neues Modell berechnet werden. Die Wald-Statistik wird daher in allen ML-Programmen automatisch mit ausgedruckt. Das gleiche gilt für die Likelihood-Verhältnis-Statistik für einige Standardvergleiche (Test des Gesamtmodells ohne Regres-

sionskonstante). Die Score−Statistik (5.33) hat schließlich den Vorteil, daß das Maximierungsproblem wegfällt, da man lediglich die Parameter der Nullhypothese in die Score−Funktionen einsetzen muß.

### *5.2.5 Anwendung der Teststatistiken an Hand des Beispiels*

Nach diesen vielen abstrakten Formeln ist es instruktiv, sie einmal an einem Beispiel anzuwenden. Für diese Berechnungen ist die Varianz−Kovarianz−Matrix **COV**($\beta$) besonders wichtig. Bevor ich also zu dem eigentlichen Beispiel komme, muß ich leider auch noch diese etwas abstrakte Größe definieren.

Die oben erwähnten 1. partiellen Ableitungen der Likelihood−Funktion können ein weiteres Mal (partiell) abgeleitet werden − und zwar genauso häufig, wie Parameter vorhanden sind. Da die beiden Gleichungen (5.24) jeweils 2 Parameter enthalten, also genau zweimal. Man erhält die 2. partiellen Ableitungen:

$$\text{(5.35a)} \qquad \frac{\partial^2 \ln L}{\partial \beta_0 \, \partial \beta_0} = -\exp \beta_0 \, \Upsilon_0 - \exp \beta_0 \exp \beta_1 \, \Upsilon_1$$

$$\text{(5.35b)} \qquad \frac{\partial^2 \ln L}{\partial \beta_0 \, \partial \beta_1} = -\exp \beta_0 \exp \beta_1 \, \Upsilon_1$$

$$\text{(5.35c)} \qquad \frac{\partial^2 \ln L}{\partial \beta_1 \, \partial \beta_0} = -\exp \beta_0 \exp \beta_1 \, \Upsilon_1$$

$$\text{(5.35d)} \qquad \frac{\partial^2 \ln L}{\partial \beta_1 \, \partial \beta_1} = -\exp \beta_0 \exp \beta_1 \, \Upsilon_1$$

Der Einfachheit halber habe ich gleich die Abkürzung $\Upsilon_j$ für die gruppenspezifischen Summen aller Wartezeiten verwendet. Diese 4 Ableitungen kann man kompakt in eine 2·2−Matrix schreiben, die man als *Hesse'sche Matrix* bezeichnet:

$$\Psi = \begin{bmatrix} -\exp \beta_0 \Upsilon_0 - \exp \beta_0 \exp \beta_1 \Upsilon_1 & -\exp \beta_0 \exp \beta_1 \Upsilon_1 \\ -\exp \beta_0 \exp \beta_1 \Upsilon_1 & -\exp \beta_0 \exp \beta_1 \Upsilon_1 \end{bmatrix}$$

Die Hesse'sche Matrix $\Psi$ wird ohnehin in dem bekanntesten Optimierungsverfahren (Newton – Raphson) zur iterativen Bestimmung des Maximums der Likelihood – Funktion benötigt. Außerdem darf sie nicht negativ definit sein, wenn Sattelpunkte (Wendepunkte) der Likelihood – Funktion ausgeschlossen werden sollen (vgl. die Anmerkung zu Gleichung 5.16). Entscheidend für die folgenden Ableitungen ist der mit $-1$ multiplizierte Erwartungswert $\mathbf{I}(\beta) = -\mathrm{E}(\Psi)$ der Hesse'schen Matrix:

$$(5.36) \qquad \mathbf{I}(\beta) = \begin{bmatrix} \exp \beta_0 \Upsilon_0 + \exp \beta_0 \exp \beta_1 \Upsilon_1 & \exp \beta_0 \exp \beta_1 \Upsilon_1 \\ \exp \beta_0 \exp \beta_1 \Upsilon_1 & \exp \beta_0 \exp \beta_1 \Upsilon_1 \end{bmatrix}$$

Diese Matrix wird auch als *Informationsmatrix* bezeichnet und ein allgemeiner Satz der ML – Theorie besagt, daß die Inverse dieser Matrix genau der Varianz – Kovarianz – Matrix der ML – Schätzer entspricht. Dieser Satz bezieht sich auf die Erwartungswerte der 2. Ableitungen, also die Werte, die man bei Kenntnis der Grundgesamtheitsparameter erwarten würde (erwartete Informationsmatrix). Da diese Parameter jedoch unbekannt sind, verwendet man in (5.36) die entsprechenden ML – Schätzer der $\beta$[1]. Man spricht dann von der beobachteten Informationsmatrix $\mathbf{I}(\hat{\beta})$ und verwendet deren Inverse als besten Schätzer für die Varianz – Kovarianz – Matrix:

$$(5.37) \qquad \hat{\mathbf{COV}}(\beta) = \mathbf{I}(\hat{\beta})^{-1}$$

$$= \begin{bmatrix} (\exp \hat{\beta}_0 \Upsilon_0)^{-1} & -(\exp \hat{\beta}_0 \Upsilon_0)^{-1} \\ -(\exp \hat{\beta}_0 \Upsilon_0)^{-1} & (\exp \hat{\beta}_0 \Upsilon_0)^{-1} + (\exp \hat{\beta}_0 \exp \hat{\beta}_1 \Upsilon_1)^{-1} \end{bmatrix}$$

Das Element in der zweiten Spalte der zweiten Zeile enthält dann die Varianz des zweiten Parameters $\beta_1$. Wenn man die entsprechenden Schätzer (5.25) für $\hat{\beta}_0$ und $\hat{\beta}_1$ einsetzt, ergibt sich folgende Varianzformel:

$$(5.38) \qquad \hat{\mathrm{VAR}}(\beta_1) = \frac{\Upsilon_0}{D_0 \Upsilon_0} + \frac{D_0 \Upsilon_1 \Upsilon_0}{D_1 \Upsilon_0 D_0 \Upsilon_1} = \frac{1}{D_0} + \frac{1}{D_1} = 0{,}0069$$

1) Dieses Vorgehen ist aus der multiplen Regression bekannt, bei der die Varianz $\sigma$ durch die Varianz s der Residuen geschätzt wird, die sich auf Grund der Modellprognosen ergeben.

mit der sich beliebige Konfidenzintervalle für den Parameter $\beta_1$ berechnen lassen. Wenn sich beide Gruppen in ihrem Abstiegsrisiko unterscheiden sollen ($H_A$: $\beta_1 \neq 0$), dann darf dieses Konfidenzintervall bei einer gegebenen Fehlerwahrscheinlichkeit den Wert Null nicht einschließen. Das 95% – Konfidenzintervall lautet für die Beispieldaten:

$$P\left[\hat{\beta}_1 - 1{,}96\sqrt{VAR(\beta_1)} \leq \beta_1 \leq \hat{\beta}_1 + 1{,}96\sqrt{VAR(\beta_1)}\right] = 0{,}95$$

$$P[-0{,}626 \leq \beta_1 \leq -0{,}300] = 0{,}95$$

Da es den Wert Null nicht enthält, kann man sagen, daß sich beide Gruppen signifikant unterscheiden (5% – Signifikanzniveau).

Eine kleine Umformung des obigen Konfidenzintervalls ergibt den sogenannten *t – Wert* für die Nullhypothese $H_0$ $\beta_1 = 0$:

$$t = \frac{|\hat{\beta}_1|}{\sqrt{VAR(\beta_1)}} = 5{,}562 \tag{5.39}$$

Der t – Wert mißt praktisch die Abweichung des Schätzwertes $\hat{\beta}_1 = -0{,}463$ vom Wert Null in Einheiten des Standardfehlers. Durch Vergleich mit einer Standardnormalverteilung kann man die Signifikanz des Parameters berechnen. Als Faustregel gilt, daß alle Parameter auf dem 5% – Niveau signifikant sind, die mindestens doppelt (genauer: 1,96 – mal) so groß sind wie ihr Standardfehler. In diesem Fall ergibt sich ein Wert von 5,56, der hoch signifikant ist. Dieser t – Wert ist jedoch nur eine Umformung des obigen Konfidenzintervalls.

Anders ist das Vorgehen beim Score – Test. Hier formuliert man zunächst wieder eine Nullhypothese, nämlich $\tilde{\beta}_1 = 0$. Unter dieser Nullhypothese schätzt man dann die verbleibenden Parameter, in diesem Fall also $\beta_0$. Durch Einsetzen von $\tilde{\beta}_1 = 0$ in die Score – Funktion (5.24a) für $\beta_0$ und erneutes Nullsetzen ergibt sich ein neuer Schätzer für die Regressionskonstante:

$$\exp \tilde{\beta}_0 = \tilde{a}_0 = \frac{D_0 + D_1}{\Upsilon_0 + \Upsilon_1} \tag{5.40}$$

Wenn sich also beide Gruppen nicht unterscheiden, dann entspricht der Antilogarithmus der Regressionskonstanten, wie zu erwarten, der Gesamt–Abstiegsrate (= Anzahl der Ereignisse insgesamt pro Jahr Berufstätigkeit insgesamt). Mit diesem Schätzer kann man den Wert des zweiten Scores (5.24b) berechnen:

$$(5.41) \quad \Theta_2(\tilde{\beta} \mid \tilde{\beta}_1 = 0) = D_1 - \Upsilon_1 \exp \tilde{\beta}_0$$

$$= D_1 - \Upsilon_1 \frac{D_0 + D_1}{\Upsilon_0 + \Upsilon_1} = \frac{D_1\Upsilon_0 - D_0\Upsilon_1}{\Upsilon_0 + \Upsilon_1}$$

Die Frage ist jetzt, ob dieser (unter der Nullhypothese sich ergebende) Score signifikant von Null verschieden ist.

Erneut muß man ein Konfidenzintervall berechnen und benötigt dafür die geschätzte Varianz von $\Theta_2$. Unter der Nullhypothese ist das Element in der 2. Zeile und 2. Spalte der Varianz–Kovarianz–Matrix (5.37) der Kehrwert der gesuchten Varianz. Man setzt den neuen Schätzer für $\beta_0$ (s. Gleichung 5.40) und den unter $H_0$ angenommenen Wert für $\beta_1$ (=0) ein und erhält die gewünschte Varianzformel:

$$(5.42) \quad \widehat{VAR}(\Theta_2) = [(\exp \tilde{\beta}_0\Upsilon_0)^{-1} + (\exp \tilde{\beta}_0 \exp \tilde{\beta}_1\Upsilon_1)^{-1}]^{-1}$$

$$= \left[\frac{\Upsilon_0 + \Upsilon_1}{(D_0 + D_1)\,\Upsilon_0} + \frac{\Upsilon_0 + \Upsilon_1}{(D_0 + D_1)\,\Upsilon_1}\right]^{-1} = \frac{(D_0 + D_1)\,\Upsilon_0\,\Upsilon_1}{(\Upsilon_0 + \Upsilon_1)^2}$$

Ähnlich (5.39) kann man auch hier eine standardnormalverteilte Teststatistik bilden:

$$(5.43) \quad t = \frac{|\Theta_2(\tilde{\beta}|\tilde{\beta}_1 = 0)|}{\sqrt{VAR(\Theta_2)}} = \frac{|D_1\Upsilon_0 - D_0\Upsilon_1|}{\sqrt{(D_0 + D_1)\,\Upsilon_0\Upsilon_1}} = 5{,}611$$

Der sich ergebende Wert von 5,61 stimmt weitgehend mit dem oben berechneten t–Wert überein und deutet erneut darauf hin, daß sich beide Gruppen signifikant in ihrem Abstiegsrisiko unterscheiden.

Eine letzte Testmöglichkeit ergibt sich durch den Vergleich der Likelihood–Funktionen der Modelle, die sich nur durch den Parameter $\beta_1$ unter-

scheiden. In Modell a habe $\beta_1$ den Wert Null (Nullhypothese) und in Modell A kann er frei variieren, wird daher durch (5.25b) geschätzt. $\beta_0$ hat dementsprechend in Modell a den Schätzwert (5.40) und in Modell A den Schätzwert (5.25a). Diese Werte können in die (logarithmierte) Likelihood–Funktion (5.23) eingesetzt werden und mit den entsprechenden Funktionswerten für Modell A und a kann man die Teststatistik (5.34) berechnen:

$$(5.44) \qquad X^2 = 2 \ln \frac{L(A)}{L(a)} = 2 [\ln L(A) - \ln L(a)]$$

$$= 2 \left\{ D_0 \ln \frac{D_0}{\Upsilon_0} + D_1 \ln \frac{D_1}{\Upsilon_1} - (D_0 + D_1) \ln \frac{D_0 + D_1}{\Upsilon_0 + \Upsilon_1} \right\} = 32{,}682$$

Es ergibt sich ein Wert von 32,68, der verglichen mit einer Chi–Quadrat–Tabelle mit einem Freiheitsgrad hoch signifikant ist. Anders ausgedrückt, die Hinzufügung eines weiteren Parameters, der die Gruppenunterschiede berücksichtigt, verbessert das Modell a mit nur einem Parameter signifikant. Der Likelihood–Verhältnis–Test ist also ein globaler Test der Modellverbesserung. Auf Grund seiner optimalen formalstatistischen Eigenschaften (vgl. Abschnitt 5.2.4) ist er besonders gut geeignet, den gemeinsamen Einfluß von mehreren Variablen zu überprüfen.

Um abschließend die Ergebnisse des Likelihood–Verhältnis–Testes mit den beiden anderen Teststatistiken zu vergleichen, greife ich auf einen allgemeinen Satz der Statistik zurück. Er besagt, daß die Wurzel aus einer Chi–Quadrat–verteilten Zufallsvariablen mit einem Freiheitsgrad normalverteilt ist. Die Wurzel aus 32,682 ergibt also eine standardnormalverteilte Prüfgröße, die mit (5.43) und (5.39) vergleichbar ist. In diesem Fall ergibt sich ein Wert von 5,72, der von den beiden anderen nur unwesentlich abweicht.

### *5.2.6 Eine Likelihood–Funktion für multiple Ereignisse*

Gleichung (5.18) beschreibt die Likelihood eines Prozesses mit singulären Ereignissen. Möchte man zwischen verschiedenen Ereignissen differenzieren, dann muß man erstens festhalten, um welches Ereignis es sich handelt (sofern eins stattfindet), und zweitens berücksichtigen, daß bei einer zensierten Beobachtung alle Risiken überlebt werden. Man kann sich diesen

Prozeß in drei Schritten vorstellen. Alle $N = 7587$ Personen beginnen zunächst im gleichen Zustand j:

1. Nach 10 Berufsjahren haben $C_0 = 3778$ Personen ihre Tätigkeit nicht verändert. Sie haben sozusagen alle Risiken überlebt.
2. Die verbleibenden $(N - C_0) = 3809$ Personen haben ihre Tätigkeit gewechselt.
3. Dieser Tätigkeitswechsel war bei $D_1 = 690$ Personen mit einem Abstieg, bei $D_2 = 2448$ Personen mit einem Aufstieg und bei $D_3 = 671$ Personen mit einer geringfügigen Veränderung des sozio-ökonomischen Status verbunden.

Dementsprechend besteht die Likelihood-Funktion aus drei Teilen:

1. aus der Überlebenswahrscheinlichkeit für $C_0$ Personen, alle drei verschiedenen Ereignisse zu überleben;
2. aus der Dichte für $(D_1 + D_2 + D_3)$ Personen, irgendeines der drei Ereignisse zu erleben; und schließlich
3. aus der bedingten Wahrscheinlichkeit für jede der drei Gruppen $D_1$ bis $D_3$, von den drei möglichen das entsprechende Ereignis zu erleben.

Seien $C$ und $D$ wieder die entsprechenden Teilmengen der Stichprobe ohne bzw. mit Wechsel (davon: $D_1$ mit Abstieg, $D_2$ mit Aufstieg, $D_3$ mit horizontaler Mobilität), dann lautet die Likelihood-Funktion entsprechend:

$$(5.45a) \qquad L = \prod_{i \in C} S_j(t_i) \prod_{i \in D} f_j(t_i) \prod_{i \in D_1} m_{j1}(t_i) \prod_{i \in D_2} m_{j2}(t_i) \prod_{i \in D_3} m_{j3}(t_i)$$

Dabei ist $S_j(t)$ die Überlebenswahrscheinlichkeit (2.26a), im Ausgangszustand j zu verbleiben, $f_j(t)$ die Dichte (2.26c) der Wartezeiten bis zu irgendeinem Wechsel und $m_{jk}(t)$ die bedingte Wahrscheinlichkeit (2.42) eines Wechsels nach k, vorausgesetzt es findet ein Wechsel statt.

Alle drei Funktionen lassen sich auf Raten zurückführen. Da jetzt multiple Ereignisse untersucht werden, benötigt man für jeden möglichen Wechsel eine Übergangsrate $r_{jk}(t)$. Der Verbleib im Zustand j wird mit Hilfe der Abgangsrate $r_{j.}(t)$ modelliert, die der Summe der einzelnen Übergangsraten entspricht (vgl. Gleichung 2.30). Für jede der Übergangsraten kann man ein spezielles Regressionsmodell formulieren und es in die Likelihood-Funktion einsetzen (vgl. Gleichung 5.2c). Diese sind dann wiederum nur eine Funktion der Zeit t sowie der Kovariaten $\mathbf{x}_{ijk}$ und der Parameter $\beta_{jk}$: $r_{jk}(t|\mathbf{x}_{ijk}, \beta_{jk})$. Dabei können Kovariaten und Parameter je nach ana-

lysiertem Übergang variieren, daher der Index j bzw. k. Mit Hilfe eines iterativen Lösungsalgorithmus kann man ML−Schätzer für die unbekannten Parameter $\beta_{jk}$ berechnen, ähnlich wie ich es im vorigen Abschnitt an Hand eines einfachen Beispiels vorgeführt habe.

Da dieses Problem in der Regel nur durch Computerprogramme gelöst werden kann, will ich mich mit diesen allgemeinen Bemerkungen begnügen. Für einige der folgenden Überlegungen ist es jedoch hilfreich, die obige Likelihood−Funktion noch etwas zu vereinfachen. Unter Verwendung einer Verallgemeinerung von (2.14) für $f_j(t)$ und (2.42) für $m_{jk}(t)$ und unter Berücksichtigung der Übergangsraten ergibt sich:

$$(5.45b)\quad L = \prod_{i\in C} S_j(t_i) \prod_{i\in D} r_{j.}(t_i) S_j(t_i) \prod_{i\in D_1} \frac{r_{j1}(t_i)}{r_{j.}(t_i)} \prod_{i\in D_2} \frac{r_{j2}(t_i)}{r_{j.}(t_i)} \prod_{i\in D_3} \frac{r_{j3}(t_i)}{r_{j.}(t_i)}$$

$$= \prod_{i\in C} S_j(t_i) \prod_{i\in D_1} r_{j1}(t_i) \prod_{i\in D_2} r_{j2}(t_i) \prod_{i\in D_3} r_{j3}(t_i)$$

Wenn man jetzt wieder eine Statusvariable definiert, die nicht nur angibt, ob ein Ereignis stattfindet, sondern auch mißt, um welchen der $k=1,\ldots,K$ Wechsel es sich dabei handelt, dann kann man die Gleichung weiter vereinfachen:

$$(5.45c)\quad L = \prod_{i}^{N} \left( S_j(t_i|\mathbf{x}_{ijk},\boldsymbol{\beta}_{jk}) \prod_{k}^{K} r_{jk}(t_i|\mathbf{x}_{ijk},\boldsymbol{\beta}_{jk})^{\delta_{ijk}} \right)$$

$$\text{mit } \delta_{ijk} = \begin{cases} 1 & \text{(Wechsel von j nach k)} \\ 0 & \text{(Zensierung)} \end{cases}$$

Schließlich machen wir noch die Annahme, daß die K Risiken unabhängig voneinander wirken. Die Überlebensfunktion $S_j(t)$ entspricht dann dem Produkt der übergangsspezifischen Pseudo−Überlebensfunktionen $S_{jk}(t)$ (vgl. Gleichung 2.29) und die Likelihood−Funktion kann in k übergangsspezifische Terme faktorisiert werden:

$$L = \prod_{k}^{K} \prod_{i}^{N} S_{jk}(t_i|\mathbf{x}_{ijk},\boldsymbol{\beta}_{jk})\, r_{jk}(t_i|\mathbf{x}_{ijk},\boldsymbol{\beta}_{jk})^{\delta_{ijk}}$$

Falls mehrere Ausgangszustände existieren, dann betrachten wir diese als gegeben und können quasi für jeden der j=1,...,J Ausgangszustände getrennt eine ML-Schätzung vornehmen. Die Gesamt-Likelihood entspricht dem Produkt über alle J Ausgangszustände:

$$L = \prod_{j}^{J} \prod_{k}^{K} \prod_{i}^{N} S_{jk}(t_i|\mathbf{x}_{ijk},\boldsymbol{\beta}_{jk})\ r_{jk}(t_i|\mathbf{x}_{ijk},\boldsymbol{\beta}_{jk})^{\delta_{ijk}} \tag{5.46}$$

(5.46) ist die gesuchte *allgemeine Likelihood-Funktion für Prozesse mit multiplen, nicht-wiederholbaren Ereignissen* (J Ausgangs- und K Zielzustände).

Die Eigenschaften dieser Likelihood-Funktion lassen sich wiederum am besten an Hand eines einfachen Beispiels diskutieren. Angenommen alle drei Arten von Tätigkeitswechseln der MZ71-Daten treten im Zeitablauf mit konstanter Rate auf. Die drei Raten seien $\lambda_{01}$, $\lambda_{02}$ und $\lambda_{03}$. Da der Ausgangszustand in diesem Beispiel keine Rolle spielt, habe ich hier für alle drei Raten eine Null eingesetzt. In der Likelihood-Funktion kann man dementsprechend die Multiplikation für j weglassen. Unter Verwendung von (2.29) ergibt sich dann:

$$L = \prod_{i}^{N} \exp(-(\lambda_{01}+\lambda_{02}+\lambda_{03})\ t_i)\ \lambda_{01}^{\delta_{i01}}\ \lambda_{02}^{\delta_{i02}}\ \lambda_{03}^{\delta_{i03}}$$

ML-Schätzer für die drei Raten findet man wiederum durch Maximierung der logarithmierten Likelihood-Funktion. Sie lautet:

$$\ln L = \sum_{i}^{N} [-(\lambda_{0i}+\lambda_{02}+\lambda_{03})\ t_i + \delta_{i01} \ln \lambda_{01} + \delta_{i02} \ln \lambda_{02} + \delta_{i03} \ln \lambda_{03}]$$

Bildet man z.B. die erste Ableitung nach $\lambda_{01}$ und setzt diese Null, erhält man folgenden Schätzer für die Abstiegsrate:

$$\frac{\partial \ln L}{\partial \lambda_{01}} = \sum_{i}^{N} \left( -t_i + \frac{\delta_{i01}}{\lambda_{01}} \right) = 0 \quad \text{mit } \hat{\lambda}_{0i} = \frac{\sum_{i}^{N} \delta_{i01}}{\sum_{i}^{N} t_i} = \frac{D_{01}}{\Upsilon}$$

Wie zu erwarten, entspricht er dem Verhältnis aller Abstiege $D_{01}$ zur Gesamtdauer aller Wartezeiten. Wichtig ist in diesem Zusammenhang jedoch, daß man dieses Ergebnis unabhängig davon erzielt, ob man die anderen Ereignisse als zensierte oder als unzensierte Beobachtungen behandelt. Im ersten Fall würden die entsprechenden Statusvariablen $\delta_{i02}$ und $\delta_{i03}$ immer den Wert Null haben. Im zweiten Fall hätten sie, wie vorgesehen, immer dann den Wert Eins, wenn ein entsprechender Wechsel stattfindet. Anders ausgedrückt, man kann einzelne *Übergangsraten unabhängig voneinander schätzen*, indem man die nicht interessierenden Ereignisse zunächst als zensierte Beobachtungen betrachtet. Dadurch werden diese konkurrierenden Ereignisse nicht gänzlich von der Analyse ausgeschlossen und tauchen als Bestandteil der Gesamtdauer aller Wartezeiten in dem obigen Schätzer wieder auf.

Von dieser Möglichkeit habe ich insbesondere bei den explorativen Verfahren Gebrauch gemacht. Mit der Likelihood (5.46) läßt sich dieses Vorgehen nachträglich legitimieren. Voraussetzung ist allerdings die Unabhängigkeitsannahme (2.29). Nur so läßt sich die Likelihood (5.45c) in k voneinander unabhängige Terme faktorisieren.

## *5.3 Interpretation von Regressionsmodellen für Verlaufsdaten*

Schwerpunkt dieses Abschnitts ist die Interpretation von Regressionsmodellen für Verlaufsdaten. An Hand der MZ71 – Daten und der in Abschnitt 3.2 diskutierten Hypothesen möchte ich zeigen, wie man (log – lineare) Effekte einzelner Variablen und die Erklärungskraft des Gesamtmodells anschaulich darstellen kann (PRG0501).

Da diese Fragen in mehr oder weniger allen Regressionsmodellen für Verlaufsdaten eine Rolle spielen, ist es dazu nicht nötig, weitergehende Modelle als das bisher diskutierte Modell exponentiell verteilter Wartezeiten einzuführen. Mit ihm lassen sich zeitkonstante, aber multivariate Raten flexibel prognostizieren. Zeitabhängige Hypothesen wie etwa der Zusammenhang zwischen Karrieremobilität und Berufserfahrung müssen zunächst ausgeklammert werden. Aus einem ähnlichen Grund werden auch zeitabhängige Kovariaten nicht berücksichtigt. Ich gehe also nicht nur von einem zeitkonstanten Prozeß, sondern auch von konstanten Rahmenbedingungen aus.

An Hand der MZ71 – Daten habe ich mich gefragt, wie häufig die verschiedenen Arten von Tätigkeitswechseln überhaupt vorkommen, unabhängig davon, daß es Unterschiede im Zeitablauf und zwischen Personen mit unterschiedlichen Eigenschaften gibt. Zu diesem Zweck habe ich das oben diskutierte Modell mit konstanten, homogenen Übergangsraten für die ersten Tätigkeiten der untersuchten Personen geschätzt (Modell 0). Es mißt sozusagen den Durchschnitt aller interpersonellen und zeitlichen Unterschiede (*Null – oder Basismodell*) und bildet den Ausgangspunkt für realistischere Modelle. Danach treten in der Stichprobe nur 0,012 Abstiege aber 0,042 Aufstiege pro Jahr Berufstätigkeit auf (vgl. die Antilogarithmen in Tabelle 5.2).

*Tabelle 5.2: Modelle mit zeitkonstanter Rate (MZ71 – Daten)*

| | | Abstiege | | | | Aufstiege | | | |
|---|---|---|---|---|---|---|---|---|---|
| Variable | | Modell 0 | Modell 1 | Modell 2 | Modell 3 | Modell 0 | Modell 1 | Modell 2 | Modell 3 |
| Konstante | a) | −4,4481* | −2,6136* | −1,6783* | −2,7030* | −3,1817* | −5,6424* | −5,6645* | −5,7443* |
| | b) | 0,0381 | 0,2671 | 0,2991 | 0,2876 | 0,0202 | 0,1111 | 0,1225 | 0,1236 |
| | c) | 0,0117 | 0,0733 | 0,1867 | 0,0670 | 0,0415 | 0,0035 | 0,0035 | 0,0032 |
| Status des Vaters | | | −0,0011 | −0,0014* | −0,0011 | | 0,0029* | 0,0026* | 0,0029* |
| | | | 0,0006 | 0,0006 | 0,0006 | | 0,0002 | 0,0002 | 0,0002 |
| | | | 0,9989 | 0,9986 | 0,9989 | | 1,0029 | 1,0026 | 1,0029 |
| Ausbildung des Befragten | | | −0,1792* | −0,1961* | −0,1789* | | 0,2824* | 0,2149* | 0,2827* |
| | | | 0,0299 | 0,0307 | 0,0299 | | 0,0117 | 0,0120 | 0,0117 |
| | | | 0,8359 | 0,8219 | 0,8362 | | 1,3263 | 1,2397 | 1,3267 |
| Status des Befragten | | | 0,0024* | 0,0046* | 0,0021* | | −0,0115* | −0,0108* | −0,0118* |
| | | | 0,0008 | 0,0009 | 0,0009 | | 0,0004 | 0,0005 | 0,0005 |
| | | | 1,0024 | 1,0046 | 1,0021 | | 0,9886 | 0,9893 | 0,9883 |
| Beschäftigten – wachstum Ausgangsbranche | | | | 0,0106* | 0,0010 | | | −0,0033* | 0,0012* |
| | | | | 0,0012 | 0,0012 | | | 0,0007 | 0,0006 |
| | | | | 1,0107 | 1,0010 | | | 0,9967 | 1,0012 |
| Beschäftigten – wachstum Zielbranche | | | | −0,0203* | | | | 0,0090* | |
| | | | | 0,0017 | | | | 0,0006 | |
| | | | | 0,9799 | | | | 1,0090 | |
| Tätigkeits – wechsel 1967 | | | | 1,2597* | | | | 0,9364* | |
| | | | | 0,1326 | | | | 0,0657 | |
| | | | | 3,5244 | | | | 2,5508 | |
| N | | 7587 | 7587 | 7587 | 7587 | 7587 | 7587 | 7587 | 7587 |
| D | | 690 | 690 | 690 | 690 | 2448 | 2448 | 2448 | 2448 |
| in % | | 9,1% | 9,1% | 9,1% | 9,1% | 32,3% | 32,3% | 32,3% | 32,3% |
| ln L | | −3759,17 | −3727,53 | −3617,76 | −3727,18 | −10[illegible]6,9 | −9727,32 | −9527,76 | −9725,58 |
| Pseudo – $R^2$ | | 0,0% | 0,8% | 3,8% | 0,9% | 0,0% | 5,0% | 6,9% | 5,0% |
| L – Ratio | | 0 | 63,28 | 282,82 | 63,97 | 0 | 1019,16 | 1418,28 | 1022,65 |
| df | | 0 | 3 | 6 | 4 | 0 | 3 | 6 | 4 |

a) Parameter (* sign. 5% – Niveau), b) Standardfehler, c) Antilogarithmus

Im nächsten Schritt wurden alle konstanten Merkmale der Person und der ausgeübten Tätigkeit berücksichtigt (Modell 1). Um den Einfluß wirtschaftlicher Rahmenbedingungen auf den Mobilitätsprozeß zu untersuchen, wurde schließlich ein weitergehendes Modell 2 geschätzt, das einige dieser Rahmenbedingungen zusätzlich beinhaltet. Die Ergebnisse dieser Berechnungen sind in Tabelle 5.2 enthalten und sollen im folgenden ausführlich interpretiert werden.

### *5.3.1 Modellfit und Modellverbesserung*

Dabei stellt sich zunächst einmal die Frage, wie gut das jeweilige Modell auf die Daten paßt. Zur Beantwortung dieser Frage verwendet man in der Regel den oben diskutierten Likelihood-Verhältnis-Test (vgl. Gleichung 5.34). Aussagen über den *Anteil erklärter Varianz* kann man aber leider nicht machen, da zwei sehr unterschiedliche Arten von Daten vorliegen. Bei Ereignissen wäre es sicherlich sinnvoll, empirische und prognostizierte Wartezeiten miteinander zu vergleichen. Bei zensierten Beobachtungen hingegen ist der Zeitpunkt des Ereignisses unbekannt. Hier könnte man lediglich prüfen, ob der Anteil zensierter Beobachtungen auch richtig vorhergesagt wird.

Der Likelihood-Verhältnis-Test untersucht, ob die Hinzufügung oder Restriktion einzelner Parameter eine signifikante Modellverbesserung bzw. -verschlechterung zur Folge hat. Man kann also prüfen, ob die beiden angeblich realistischeren Modelle 1 und 2 wirklich besser als das einfache Basismodell sind. Dazu bildet man die doppelte Differenz der entsprechenden maximierten (logarithmierten) Likelihood-Funktionen und vergleicht sie mit einer Chi-Quadrat-Verteilung. Bei den Abstiegen ergibt sich z.B. ein Wert von 63,3, wenn man Modell 1 mit Modell 0 vergleicht:

$$2\,[(-3727{,}53) - (-3759{,}17)] = 63{,}28.$$

Da Modell 1 zusätzlich 3 Parameter schätzt, muß dieser Wert mit einer Chi-Quadrat-Verteilung mit 3 Freiheitsgraden verglichen werden. Danach ist dieser Wert hoch signifikant.

Natürlich kann man auch prüfen, ob die Berücksichtigung der wirtschaftlichen Rahmenbedingungen gegenüber den schon im Modell befindlichen Merkmalen einen zusätzlichen Informationsgewinn bedeutet. Hierzu

vergleicht man Modell 2 und 1 und auch hier ist die entsprechende Teststatistik hoch signifikant (219,5 mit 3 Freiheitsgraden).

Verschiedentlich wird der Wert der Likelihood−Funktion skaliert, um eine Maßzahl zu erhalten, die der Zahl R−Quadrat ähnlich ist:

$$\text{Pseudo}-R^2 = 1 - \frac{\ln L \text{ (aktuelles Modell)}}{\ln L \text{ (Vergleichsmodell)}} \tag{5.47}$$

Dieses *Pseudo−R−Quadrat* mißt die relative Verringerung der Likelihood−Funktion des aktuellen Modells gegenüber einem geeigneten Vergleichsmodell (meistens das Modell 0 mit zeitkonstanter, homogener Rate). Häufig wird diese Zahl wie der Anteil erklärter Varianz interpretiert. Danach würde das Modell 1 0,8% und das Modell 2 3,8% der Varianz der Abstiege erklären. Wenn diese Interpretation stimmen würde, wäre das ein sehr schlechter Modellfit.

In der Tat handelt es sich dabei um eine Überinterpretation. Dazu sollte man sich einmal überlegen, was sich hinter der Maßzahl (5.47), genauer gesagt hinter den Likelihood−Funktionen verbirgt. Eine Likelihood−Funktion ist das Produkt vieler Einzeldichten, d.h. vieler kleiner Werte. Sie wird daher ebenfalls einen sehr kleinen Wert haben, der um so kleiner wird, je mehr Dichten man miteinander multipliziert, je mehr Fälle man also untersucht. Durch abschließende Logarithmierung erhält man dann eine große negative Zahl (vgl. Tabelle 5.2). Je mehr Fälle man also betrachtet, um so größer wird der Nenner in (5.47) und um so kleiner die relative Modellverbesserung. Wird diese dann fälschlicherweise als erklärte Varianz interpretiert, suggeriert sie in der Tat einen sehr schlechten Modellfit.

Man sollte daher nur den Begriff relative Modellverbesserung verwenden. Von diesem Standpunkt aus gesehen, tritt bei den Abstiegen die entscheidende Modellverbesserung mit der Berücksichtigung wirtschaftlicher Rahmenbedingungen ein, während bei den Aufstiegen das Modell nicht wesentlich verbessert wird. Hier stellt schon Modell 1 eine wesentliche Verbesserung gegenüber dem Basismodell dar, d.h. persönliche und tätigkeitsspezifische Merkmale tragen sehr viel mehr zur Erklärung beruflicher Aufstiege bei. Bei diesen einfachen zeitkonstanten Modellen zeigen sich also schon verschiedene Muster intragenerationeller Mobilität wie in Abschnitt 3.2 vermutet.

Abschließend sei noch darauf hingewiesen, daß bisher nur die relative Modellverbesserung gegenüber einem geeigneten Basismodell betrachtet

wurde. Wie gut das Modell in absoluten Zahlen ist, weiß man dann noch immer nicht. An dieser Stelle vermißt man eine Maßzahl, die es erlaubt, den Modellfit in absoluten Zahlen auszudrücken. Das ist jedoch ein genereller Nachteil von ML−Schätzungen. Sie berechnen die Schätzer, die unter den Modellannahmen und gegebenen Daten die plausibelsten sind. Das Modell 0 mit zeitkonstanter, homogener Rate ist z.B. ein sehr unrealistisches Modell. Dennoch ist unter dieser Annahme der jeweilige Schätzer der Rate der plausibelste.

### *5.3.2 Signifikanz und Richtung der einzelnen Effekte*

Der nächste Untersuchungsschritt ist eine Überprüfung der einzelnen Variableneffekte. Am einfachsten läßt sich die Frage entscheiden, welche Parameter signifikant von Null verschieden sind und in welche Richtung die Effekte im einzelnen wirken.

Im ersten Fall muß man lediglich prüfen, ob der jeweilige Parameter in etwa doppelt so groß ist wie sein Standardfehler (vgl. Gleichung 5.39). Über die zweite Frage entscheidet das Vorzeichen des Parameters. Alle notwendigen Zahlen findet man dazu ebenfalls in Tabelle 5.2. Es zeigt sich, daß mehr oder weniger alle Effekte der Modelle 1 und 2 auf dem 5% Niveau signifikant von Null verschieden sind (sie sind durch ein * gekennzeichnet). Das ist in diesem Fall sicherlich auch ein Resultat der großen Fallzahl, denn bei großen Stichproben sind auch die kleinsten Unterschiede signifikant.

Ein zweiter Blick auf die Vorzeichen der einzelnen Parameter zeigt darüber hinaus eine bemerkenswerte Konstanz der Effekte über alle Modelle. Ein negatives Vorzeichen bedeutet, daß die entsprechende Variable das Abstiegsrisiko (die Aufstiegschance) verringert. Ein positives Vorzeichen signalisiert den gegenteiligen Effekt. Vergleicht man danach die Schätzungen mit den in Tabelle 3.3 erwarteten Effekten, kann man feststellen, daß die meisten Hypothesen richtig waren. Auf einige Ausnahmen werde ich weiter unten eingehen.

### *5.3.3 Prognosen*

Wie kann man jedoch die konkreten Parameterschätzungen interpretieren? Was sagt einem z.B. ein Wert von −0,1961 über den konkreten Einfluß des Merkmals Qualifikation auf das Abstiegsrisiko (Modell 2)? Eigentlich ist diese Zahl doch relativ unanschaulich. Schon bei der Frage nach dem relativen Einfluß der einzelnen Variablen (welche hat z.B. den größten Einfluß?) wird es schwierig: Bedeutet etwa ein Parameter von 1,26, daß die entsprechende Variable "Tätigkeitswechsel 1967" mehr Einfluß als das Merkmal Qualifikation hat, bei dem der entsprechende Parameter nur den Betrag 0,1961 hat? Da es sich um unstandardisierte Regressionskoeffizienten handelt, ist die Frage so einfach nicht zu beantworten. Ein Merkmal, das über einen sehr viel größeren Wertebereich streut, hat einen anderen Regressionskoeffizienten als ein Merkmal, das nur die Werte 0 und 1 annehmen kann.

Zunächst kann man die *unstandardisierten Regressionskoeffizienten* für Prognosen verwenden. Man berechnet das geschätzte Abstiegsrisiko für eine beliebige Person, indem man ihre Merkmale in die Regressionsgleichung einsetzt. Beispielsweise ergibt sich für einen Facharbeiter im verarbeitenden Gewerbe (Status = 60) mit einer gewerblichen Lehre (Qualifikation = 10), dessen Vater Meister (Status = 103) war, in Modell 1 ein geschätztes Abstiegsrisiko von 0,0126:

$$\exp(-2{,}614 - 0{,}0011 \cdot 103 - 0{,}1792 \cdot 10 + 0{,}0024 \cdot 60) = 0{,}0126.$$

Dieses Abstiegsrisiko liegt leicht über dem Durchschnitt von 0,0117 (vgl. Modell 0). Wem diese Zahl zu unanschaulich ist, der kann mit dieser Rate beliebige andere Charakteristika eines sozialen Prozesses berechnen (vgl. Abschnitt 2.2.5). Hier nur einige Beispiele:

- Eine (zeitkonstante) Abstiegsrate von 0,0126 besagt zunächst, daß in dieser Personengruppe 0,0126 Abstiege pro Jahr Berufstätigkeit auftreten.
- Weiterhin ergibt sich, daß es durchschnittlich fast 80 Jahre dauert, bis in dieser Gruppe ein Abstieg auftritt.
- Schließlich kann man prognostizieren, daß innerhalb von 10 Berufsjahren nur 11,8% dieser Gruppe einen Abstieg erfahren werden.

Das ist nur eine kurze Darstellung von möglichen Interpretationen der prognostizierten Übergangsrate. Z.B. kann man durch eine Gegenüberstel-

lung von besonders interessierenden Kontrastgruppen den Einfluß einzelner Variablen verdeutlichen. Der Phantasie des Anwenders sind dabei keine Grenzen gesetzt. Allerdings muß man dabei immer den Umweg über Prognosen für einzelne Subgruppen der Stichprobe gehen. Das erfordert zusätzliche Berechnungen und ist vor allen Dingen immer an die ausgewählten Gruppen gebunden. Es stellt sich daher die Frage, ob man die Regressionsparameter nicht auf eine einfachere und direktere Art und Weise veranschaulichen kann.

Bevor ich mich mit dieser Frage beschäftige, möchte ich nur noch ein paar Anmerkungen zu dem o.g. *Durchschnittsrisiko* $\bar{r}$ machen. Es hat für diese anschauliche Interpretation einen zentralen Stellenwert, weil ich es als Vergleichsmaßstab verwenden möchte. Formal ergibt es sich, wenn man den Mittelwert jeder Variablen (vgl. Tabelle 3.3) in die Regressionsgleichung einsetzt. Einfacher ist es jedoch, gleich die Übergangsrate des zeitkonstanten, homogenen Regressionsmodells 0 zu verwenden. Wie sich leicht nachrechnen läßt, ergibt sich bei Verwendung der Mittelwerte genau der gleiche Wert. Die Verwendung von Mittelwerten macht jedoch nur bei metrischen Merkmalen einen Sinn, nicht aber bei 0/1 – kodierten Variablen (Dummies), mit denen man nicht – metrische Merkmale (Gruppen) erfaßt. Anders ausgedrückt, die Abstiegsrate von 0,0117 ist auch der Durchschnitt der beiden Gruppen, die sich durch das Merkmal "Tätigkeitswechsel 1967" ergeben. Da eine solche Durchschnittsgruppe ein wenig anschauliches Konstrukt ist, empfehle ich, eine der Gruppen als Vergleichsgruppe zu bestimmen und für diese Gruppe das Durchschnittsrisiko neu zu berechnen. Vergleichsgruppe sollen diejenigen Personen sein, bei denen alle Dummy-Variablen den Wert 0 haben. In diesem Fall sind das die Personen, deren Tätigkeit nicht 1967 endet. Das durchschnittliche Abstiegsrisiko beträgt hier $\bar{r} = 0{,}0110$ (Aufstiege: $\bar{r} = 0{,}0397$). Wie man sich leicht überlegen kann, erhält man dieses Durchschnittsrisiko, indem man von der Regressionskonstanten des Modells 0 den mit dem entsprechenden Regressionsparameter multiplizierten Mittelwert jeder Dummy – Variablen abzieht und dann die Rate neu berechnet. Z.B. ergibt sich bei den Abstiegen: $\bar{r} = \exp(-4{,}4481 - 1{,}26 \cdot 0{,}04877) = 0{,}0110$.

### 5.3.4 *Anschauliche Interpretation log−linearer Effekte*

Log−lineare Effekte haben zwei Eigenschaften, die ihre Interpretation besonders erschweren:

1. Der Effekt einer Kovariaten variiert mit dem Wert der Variablen selbst.
2. Der Effekt einer Kovariaten variiert nicht nur mit ihrem eigenen Wert, sondern auch mit den Werten aller anderen Variablen.

Man kann sich diese Eigenschaften formal ableiten, am besten werden sie jedoch an einem Beispiel deutlich.

Für die anschauliche Interpretation verwendet man am besten die *Antilogarithmen* a (vgl. Tabelle 5.2). Wie im vorherigen Abschnitt 5.2 schon kurz erwähnt, messen sie den Faktor, mit dem man die Basisrate (dort die Rate der Personen mit niedriger Qualifikation) multiplizieren muß, wenn man die Kovariate um eine Einheit erhöht (dort die Dummy−Variable Qualifikation). Antilogarithmen haben also den Wert 1, wenn eine Variable keinen Effekt hat, und einen Wert kleiner (größer) als 1, wenn eine Variable einen negativen (positiven) Effekt hat. Dieser multiplikative Einfluß läßt sich am besten in Prozenten ausdrücken: Wenn sich das Merkmal x um eine Einheit erhöht, dann verändert sich die Rate um $100 \cdot (a-1)$ Prozent. Wenn man also die Ausbildung in Modell 2 um 1 Jahr verlängern würde, dann würde sich das Abstiegsrisiko um 17,8% verringern. Bei 0/1−kodierten Variablen (Dummies) entspricht diese Veränderung genau dem Unterschied zwischen der ($x=0$) Vergleichsgruppe und der Gruppe mit der untersuchten Eigenschaft ($x=1$). Bei Tätigkeitswechseln im Jahr 1967 ist also das Abstiegsrisiko um 252,5% größer als bei Wechseln in anderen Jahren.

Eine Erhöhung um eine Einheit hat aber bei den einzelnen Variablen auf Grund unterschiedlicher Maßeinheiten eine ganz unterschiedliche Bedeutung. Bei der Interpretation der Regressionsparameter mit Hilfe der Antilogarithmen sollte man daher gleich die unterschiedlichen Maßeinheiten der Kovariaten berücksichtigen, um auch Aussagen über den relativen Einfluß der einzelnen Variablen machen zu können. Ich schlage daher vor, zu Vergleichszwecken das jeweilige Merkmal um eine Standardabweichung zu erhöhen. Die relative Änderung der Übergangsrate bei Erhöhung einer beliebigen Variablen x (alle anderen Variablen werden als konstant betrachtet) um $\Delta x_p$ Einheiten erhält man dann durch folgende Formel:

(5.48) $$100\ (a^{\Delta x} - 1)$$

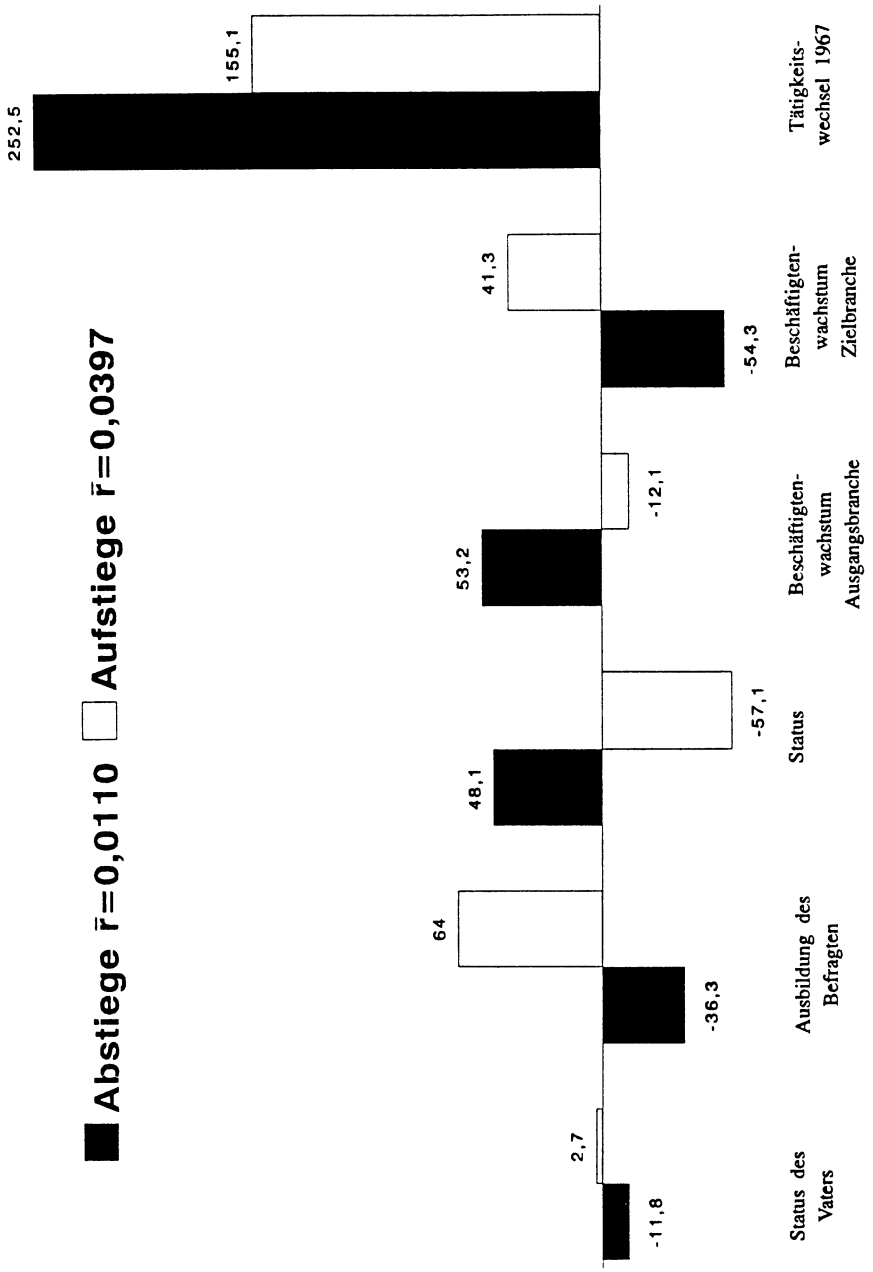
Für die Variable "Qualifikation" mit der Standardabweichung $s = 2{,}299$ ergibt sich bspw. eine Veränderung um $100 \cdot (0{,}822^{2{,}299} - 1) = -36{,}3\%$. Eine Erhöhung um eine Standardabweichung macht natürlich nur für metrische Merkmale einen Sinn. Bei 0/1-kodierten Dummies kann sich dagegen die Variable prinzipiell nur um eine Einheit erhöhen. Hier bleibt nur die Möglichkeit, die beiden Gruppen miteinander zu vergleichen (s. oben). Mit der Formel (5.48) und den Antilogarithmen aus Tabelle 5.2 habe ich die solchermaßen "standardisierten" Effekte des Modells 2 berechnet und in der folgenden Abbildung 5.1 graphisch dargestellt.

Aus der Abbildung kann man die Durchschnittsrate für die Vergleichsgruppe (Personen ohne Wechsel 1967) entnehmen – getrennt nach Art des Tätigkeitswechsels. Gleichzeitig wird deutlich, wie stark die einzelnen Variablen diese Durchschnittsrate verändern, wenn man sie um eine vergleichbare Einheit verändert. Aus den eingangs geschilderten Gründen kann man jedoch leider mit diesen Zahlen nicht so umgehen, wie man es von linear-additiven Regressionsmodellen gewöhnt ist. Beispielsweise verringert eine längere Ausbildung das Abstiegsrisiko um 36,3%. Eine weitere Verlängerung der Ausbildung (Erhöhung um 2 statt um 1 Standardabweichungen) verringert das Abstiegsrisiko jedoch nicht um $2 \cdot 36{,}3\% = 72{,}6\%$ sondern nur um 59,4%. Mit anderen Worten, der Einfluß der Qualifikation nimmt mit zunehmender Dauer ab bzw. variiert mit ihrem eigenen Wert (*nicht-linearer Effekt*). Wie man sich an Hand der Formel (5.48) überlegen kann, liegt das daran, daß die Variablenänderung $\Delta x$ jetzt im Exponenten steht. Bei einem Antilogarithmus, der kleiner als 1 ist (negativer Effekt), nimmt daher der Effekt der Variablen mit steigender Rate ab. Bei einem Antilogarithmus, der größer als 1 ist (positiver Effekt), kehrt sich diese Beziehung genau um: Der Effekt nimmt mit steigender Rate zu.

Die nächste Frage lautet: Was passiert, wenn sich zwei Merkmale gleichzeitig ändern, also z.B. die Ausbildung verlängert wird und ein Tätigkeitswechsel im Krisenjahr 1967 stattfindet? Verändert sich jetzt das Abstiegsrisiko um 252,5% – 36,3% = 216,2%? Natürlich nein, denn beide *Effekte wirken multiplikativ*. Man muß daher beide Antilogarithmen mit der jeweiligen Veränderung $\Delta x$ des Merkmals potenzieren und dann das Ergebnis miteinander multiplizieren:

$$100\ (0{,}822^{2{,}299} \cdot 3{,}525^{1} - 1\ ) = 124{,}6$$

*Abbildung 5.1: Prozentuale Veränderung der Durchschnittsrate $\bar{r}$ der Vergleichsgruppe (Personen ohne Wechsel 1967)*



In diesem Fall würde sich also das Abstiegsrisiko um 124,6% erhöhen. Mit anderen Worten, der Effekt einer Variablen ist nicht unabhängig von den Effekten anderer Variablen, so daß man die jeweiligen Effekte addieren könnte. Daher gelten die aus Abbildung 5.1 ersichtlichen Veränderungen nur unter Konstanthaltung der verbleibenden Variablen (*ceteris–paribus–Klausel*)[1].

Schließlich sollte bei prozentualen Veränderungen berücksichtigt werden, daß sie immer von der Ausgangsbasis abhängen. Eine 5–prozentige Tariferhöhung zahlt sich für die unteren Einkommensklassen ja auch sehr viel weniger aus als für die höheren Einkommensklassen. Dementsprechend verringert eine verlängerte Ausbildung das Abstiegsrisiko der Vergleichsgruppe um einen Betrag von 0,004 Einheiten (also 0,004 Abstiege weniger pro Jahr Berufserfahrung), für jede andere Gruppe ergibt sich jedoch eine ganz unterschiedliche absolute Veränderung. Mit anderen Worten, je nachdem welchen Vergleichsmaßstab man wählt (sprich: welchen Wert die anderen Merkmale haben), fällt die absolute Änderung ganz unterschiedlich aus.

Trotz dieser Interpretationsprobleme gibt die Abbildung die relativen Größenverhältnisse angemessen wieder. Aus ihr wird deutlich, wie die verschiedenen Tätigkeitswechsel auf unterschiedliche Art und Weise von den Kovariaten abhängen. Aufstiegschancen hängen im wesentlichen von der Qualifikation und dem aktuellen Status ab, während das Abstiegsrisiko in weitaus stärkerem Maße von den wirtschaftlichen Rahmenbedingungen beeinflußt wird (Beschäftigtenwachstum). Einen unerwartet hohen Einfluß, der alle anderen übertrifft, hat sowohl bei Ab– wie bei Aufstiegen das Merkmal "Tätigkeitswechsel 1967". Dieser Einfluß widerspricht auch meinen Erwartungen (vgl. Tabelle 3.3), da ich davon ausgegangen bin, daß im Krisenjahr Abstiege sehr viel wahrscheinlicher sind, Aufstiege aber sehr viel weniger auftreten. Wie erklärt sich dieser Widerspruch?

Eigentlich ist diese Variable gar nicht zeitkonstant. Sie ändert ihren Wert nur, wenn im Jahr 1967 ein Ereignis eintritt. Anders ausgedrückt, ihr Wert ist in dieser Kodierung von den Ergebnissen des untersuchten Prozesses (Eintritt eines Ereignisses) abhängig. Ein Wert von 1 bei diesem Merkmal bedeutet automatisch, daß ein Tätigkeitswechsel stattfand. Die Variable "Tätigkeitswechsel 1967" und die Zustandvariable SF1 korrelieren perfekt

1) Da die Kovariaten miteinander korrelieren (längere Ausbildung ist z.B. mit höheren Statuspositionen verbunden), ist die ceteris – paribus – Annahme nicht besonders realistisch.

miteinander. Ein ähnliches Argument trifft auf das Beschäftigtenwachstum in der Zielbranche zu. Dieses ist ja nur bekannt, wenn tatsächlich ein Wechsel stattfand. Seine Werte sind also auch vom Verlauf des untersuchten Prozesses abhängig. Um hier keine fehlenden Werte zu produzieren, wurde immer der Wert der Ausgangsbranche verwendet, wenn kein Tätigkeitswechsel stattfand (zensierte Beobachtung). Hier korrelieren also zwei Kovariaten perfekt miteinander, wenn eine zensierte Beobachtung vorliegt.

Modell 2 ist offensichtlich fehlspezifiziert.[1] Ich habe es hier nicht nur aus historischen Gründen übernommen, sondern auch, weil man aus Fehlern am besten lernt. Wir stellen also fest, daß man Variablen, deren Werte von den Ergebnissen des Prozesses abhängig sind (sog. interne Kovariaten, vgl. Abschnitt 6.1), tunlichst nicht als Kovariate in einem Regressionsmodell verwenden sollte. Den Einfluß des Krisenjahres 1967 überprüft man besser durch ein zeitabhängiges Modell, in dem die Abstiegsrate in diesem Jahr einen höheren Wert haben kann als in allen anderen Jahren des Untersuchungszeitraums (vgl. Abschnitt 5.4.2). Den Einfluß der Beschäftigungsentwicklung untersucht man besser durch eine Neudefinition des Zustandsraumes. Ähnlich wie bei der Unterscheidung zwischen statusmäßig höheren und niedrigeren Tätigkeiten für die Zustandsvariable SF1 müßte man dazu zwischen über- und unterdurchschnittlich wachsenden Wirtschaftszweigen unterscheiden und eine neue Zustandsvariable SF2 bilden.

Beide Erweiterungen sollen an dieser Stelle nicht weiter verfolgt werden. Stattdessen habe ich ein reduziertes Modell 3 geschätzt, das die beiden problematischen Variablen nicht mehr enthält (s. Tabelle 5.2). Dieses Modell ist nicht wesentlich besser als Modell 1 – zumindest bei den Abstiegen, wo der partielle Likelihood-Verhältnis-Test einen Wert von 63,97−63,28=0,69 bei einem Freiheitsgrad hat (zum Vergleich Aufstiege: 3,49). Gleichzeitig hat sich das Vorzeichen der Variablen "Beschäftigtenwachstum Ausgangsbranche" verändert, was erneut ein Zeichen dafür ist, daß beide Beschäftigtenindizes aus den o.g. Gründen hochgradig kollinear sind, so daß der Effekt der einen sich ändert, wenn die andere aus dem Modell genommen wird.

Von unseren inhaltlichen Schlußfolgerungen (s. oben) lassen sich also nur wenige Aussagen halten: Erstens können wir soziale Aufstiege sehr viel

1) Unerwartet hohe Schätzwerte eines Regressionskoeffizienten, wie etwa bei "Täigkeitswechsel 1967" in Modell 2, sind häufig ein Hinweis darauf, daß das Modell fehlspezifiziert ist.

besser erklären als Abstiege. Von den wirtschaftlichen Rahmenbedingungen ist zweitens nur eine Variable übriggeblieben. Sie hat im Prinzip keinen Einfluß auf das Abstiegsrisiko. Für soziale Aufstiege zeigt sich dagegen, daß diese um so wahrscheinlicher sind, je stärker die Beschäftigtenzahlen in dem Wirtschaftszweig wachsen, in dem die Tätigkeit ausgeübt wird, aus der die Person wechselt. Es ist zu vermuten, daß dieser Wechsel innerhalb desselben überproportional wachsenden Wirtschaftszweiges stattfindet, wenn weiterhin von einem Pull-Effekt von Arbeitsplätzen ausgegangen werden soll. Um diese Frage zu prüfen, müßte man Veränderungen des sozialen Status, die Variable SF1, mit der neuen Zustandsvariablen SF2 kreuztabellieren.

## *5.4 Zeitabhängige Modelle*

Das bisher diskutierte Modell exponentiell verteilter Wartezeiten geht von einem konstanten Ereignisrisiko aus. In diesem Abschnitt sollen Modelle diskutiert werden, die es erlauben, Veränderungen des Prozesses im Zeitablauf zu betrachten. Eine naheliegende Lösung wäre, die Variable "Zeit" auf irgendeine Art und Weise in der Regressionsgleichung zu berücksichtigen. Angenommen wir tauschen in dem einfachen Modell (5.11) die Variable "Qualifikation" $x_{i1}$ durch eine Variable t aus, die die Zeit seit Eintritt in das Berufsleben messen soll ("Berufserfahrung", vgl. Tabelle 3.3):

$$(5.49) \qquad r(t) = \exp(\beta_0 + \beta_1 t) = \exp(\beta_1 t) \exp \beta_0 = \lambda_0(t)$$

Die Basisrate $\lambda_0(t)$ ist jetzt keine Konstante mehr, sondern eine Exponentialfunktion der Zeit: $\lambda_0(t) = \exp(\beta_0 + \beta_1 t)$. Damit ist die Grundannahme exponentialverteilter Wartezeiten (zeitkonstante Rate) nicht mehr gegeben und die Formeln für Dichte f(t) bzw. Überlebensfunktion S(t) in der Likelihood (5.21) sind dementsprechend falsch. Tatsächlich entspricht die obige Regressionsgleichung (5.49) der Rate einer Gompertz-Verteilung (vgl. Anhang B und setze $\lambda = \exp \beta_0$ sowie $\gamma = \beta_1$). Hätte man hingegen den natürlichen Logarithmus ln t der Berufserfahrung verwendet

$$(5.50) \qquad r(t) = \exp(\beta_0 + \beta_1 \ln t) = t^{\beta_1} \exp \beta_0 = \lambda_0(t)$$

ergäbe sich die Rate der Weibull-Verteilung (vgl. Anhang B und setze $\gamma\lambda^{\gamma} = \exp \beta_0$ sowie $\gamma = \beta_1 + 1$). Ganz allgemein kann man daher sagen, die

Verwendung zeitabhängiger Kovariaten, die selbst eine Funktion der Prozeßzeit sind, resultiert – je nach funktionaler Abhängigkeit – in einem bestimmten Verteilungsmodell für Wartezeiten mit zeitabhängiger Rate.

Auch wenn man keine zeitabhängigen Kovariaten berücksichtigt, wird man häufig die Annahme exponentiell verteilter Wartezeiten für restriktiv halten. Je nach Fragestellung wird man ggfs. Veränderungen des Ereignisrisikos im Zeitablauf unterstellen und dementsprechend ein Verteilungsmodell mit zeitabhängiger Rate unterstellen. Bei der Fülle der Verteilungsmodelle stellt sich jedoch die Frage, nach welchen Kriterien man eigentlich eine Auswahl treffen sollte.

- Zunächst wäre zu prüfen, ob sich hinter der jeweiligen Verteilung ein stochastisches Modell verbirgt, das den untersuchten Prozeß angemessen beschreibt. Wenn z.B. Abstiege zufällig auftreten, dann ist ein Modell mit zufälligem Ereignisrisiko angemessen (Exponentialverteilung).
- Eine ähnliche Frage wäre, ob das Verteilungsmodell wünschenswerte Eigenschaften besitzt, die auf den untersuchten Prozeß zutreffen. Beispiele wären etwa ein Zeitpunkt, vor dem keine Ereignisse auftreten, oder Personen, die sich nicht verändern.
- Man kann alle Verteilungsmodelle auch nur unter dem Blickwinkel betrachten, inwieweit sie eine Verallgemeinerung der Exponentialverteilung ermöglichen. Da man dieses Modell einfach anwenden und interpretieren kann, ginge es also nur darum, es so flexibel zu machen, daß es auch Veränderungen im Zeitablauf und dabei unterschiedliche Verläufe der Rate berücksichtigt.
- Schließlich kann man ein allgemeines Modell der Rate formulieren, das alle anderen Verteilungsmodelle als Spezialfälle einschließt. Innerhalb dieses allgemeinen Rahmens kann man dann verschiedene Modelle gegeneinander testen.

Einige dieser Kriterien möchte ich in Abschnitt 5.4.1 ausführlicher diskutieren. Nach diesem allgemeinen Überblick interessiert dann wieder die Frage, wie man ein solches zeitabhängiges Modell (unter Berücksichtigung der beobachteten Heterogenität) mit Hilfe empirischer Daten schätzt. Im Prinzip ändert sich nichts an dem bisherigen Vorgehen, insbesondere ist es nicht notwendig, die bisher verwendete Likelihood–Funktion abzuändern. Das konkrete Vorgehen werde ich wieder an Hand der MZ71–Daten in Abschnitt 5.4.2 illustrieren.

Die hier vorzustellenden Verteilungsmodelle haben den Vorteil, daß sich alle Charakteristika eines sozialen Prozesses vorhersagen lassen, vorausgesetzt das Modell ist richtig. Das impliziert insbesondere einen bestimmten zeitlichen Verlauf der Rate. Häufig ist jedoch die Veränderung des Ereignisrisikos nicht bekannt oder so unregelmäßig, daß es nicht durch eine stetige Funktion approximiert werden kann. In diesem Fall ist es sinnvoller, keine spezielle Aussage über den zeitlichen Verlauf zu machen und stattdessen nur die Veränderungen im Zeitablauf zu kontrollieren. Das Regressionsmodell von COX (1972) hat diese Eigenschaft, dafür aber den Nachteil, daß ein Teil des Modells nicht spezifiziert ist. Dieses partiell parametrische Modell erfordert daher spezielle Methoden und wird im folgenden Abschnitt 5.5 besprochen.

### *5.4.1 Auswahlkriterien*

Zuerst möchte ich drei, sich teilweise ergänzende Kriterien diskutieren, ein geeignetes zeitabhängiges Modell für den untersuchten Prozeß auszuwählen. Da es immer sinnvoll ist, theoretische Gründe für ein bestimmtes Modell zu haben, möchte ich mit diesem Auswahlkriterium beginnen.

#### 5.4.1.1 Stochastisches Modell

Häufig formuliert man bestimmte Gesetzmäßigkeiten, die den untersuchten Prozeß beschreiben sollen. GOMPERTZ (1825) hat z.B. in seinem Mortalitätsgesetz behauptet, daß die menschliche Widerstandskraft mit einer Geschwindigkeit abnimmt, die dieser Widerstandskraft proportional ist. Sei w(t) die menschliche Widerstandskraft, dann besagt das Gesetz also, daß ihre Veränderung dw(t)/dt ein konstantes Vielfaches $\gamma$ von w(t) ist:

$$\frac{dw(t)}{dt} = \gamma w(t) \tag{5.51}$$

Die Mortalitätsrate r(t) steht im umgekehrten Verhältnis zur Widerstandskraft. Setzt man $w(t)=r(t)^{-1}$ in (5.51) ein, ergibt sich:

$$\frac{d\, r(t)^{-1}}{dt} = -\gamma\, r(t)^{-1}$$

Die Lösung dieser Differentialgleichung lautet (vgl. mathematischer Anhang):

$$\ln (r(t)^{-1}) = -\gamma t + \varkappa$$

und nach einigen Umformungen ergibt sich die mittlerweile bekannte *Gompertz-Rate* (vgl. Abbildung B.8):

(5.52) $\quad r(t) = \exp(\gamma t - \varkappa) = \lambda \exp(\gamma t) \quad \text{mit } \lambda = \exp(-\varkappa)$

MAKEHAM (1860) hat zusätzlich vorgeschlagen, zufällig auftretende Unfälle durch Addition einer Konstanten $\lambda_0$ in (5.52) zu berücksichtigen:

(5.53) $$r(t) = \lambda_0 + \lambda_1 \exp(\gamma t)$$

Dieses *Gompertz-Makeham-Gesetz* hat sich bei Untersuchungen der Mortalität (Personen ab 20 Jahre) bewährt und wird daher relativ häufig in der Demographie angewendet.

Die anderen Verteilungsmodelle lassen sich aus ähnlichen stochastischen Modellen ableiten. Die *Extremwertverteilung* (genauer gesagt: die Extremwertverteilung für Minima) ergibt sich bei Betrachtung der Minima von Zufallsvariablen. Untersucht man also die Lebensdauer einer Maschine, die aus mehreren Teilen besteht und die alle für das Funktionieren wichtig sind, dann hängt die Lebensdauer der Maschine von der (minimalen) Lebensdauer des Teils ab, das zuerst ausfällt. In diesem Fall würde sich also die Extremwertverteilung als stochastisches Modell anbieten. Anders hingegen, wenn alle Teile ausfallen müssen, bis die Maschine verschrottet wird. In diesem Fall könnte man die *Gamma-Verteilung* verwenden, da sie die Summe zufällig (exponentiell) auftretender Ereignisse beschreibt. ELANDT-JOHNSON/JOHNSON (1980: 414ff.) diskutieren verschiedene dieser Modelle in bezug auf eine Beschreibung von Altersprozessen und chronischen Krankheiten. DIEKMANN (1990) zeigt, daß verschiedene Prozesse sozialer Diffusion einen bestimmten Verlauf der Rate implizieren.

Nicht immer ist es jedoch möglich, den komplexen Verlauf eines Prozesses in Form eines Gesetzes zusammenzufassen. Gleichwohl hat man bestimmte Vorstellungen über einzelne Eigenschaften des Prozesses, z.B.

- über die Art der Veränderung (zunehmendes, abnehmendes, wechselndes Risiko);

- über die Geschwindigkeit, mit der Veränderungen auftreten (gleichmäßig, zunehmend, abnehmend);
- über den Zeitpunkt von Ereignissen (Wann sind Ereignisse das erste Mal möglich?);
- über die Betroffenheit von Ereignissen (Gibt es Personen ohne Ereignis oder scheidet ein Teil der Stichprobe schon gleich zu Beginn des Prozesses aus?).

Untersuchen wir einmal unter diesen Gesichtspunkten das Anwendungsbeispiel "Karrieremobilität".

Nach den Überlegungen in Abschnitt 3.2 soll die Aufstiegsrate zunehmen, während die Abstiegsrate mehr oder weniger konstant ist. Abstiege wird man daher mit der Exponentialverteilung modellieren oder man wird eine Verallgemeinerung verwenden, um diese Annahme zu testen. Für die Analyse der Aufstiege kommt dagegen nur eine Verteilung mit einer zunehmenden Rate in Frage, also z.B. die *Weibull*– oder die *Gompertz*–*Verteilung*. Beide Verteilungen unterscheiden sich nur durch die Geschwindigkeit, mit der sich r(t) ändert. Das Risiko wird langfristig bei der Gompertz–Rate sehr viel schneller zunehmen, weil hier die Zeit im Exponenten steht (exponentielles Wachstum). Da die Hypothesen keine weiteren Anhaltspunkte über die Geschwindigkeit der Veränderung liefern, ist es eine Frage persönlicher Präferenzen oder der Verfügbarkeit entsprechender Software, welche Verteilung man verwendet.

Ähnliche Überlegungen stellen DIEKMANN/MITTER (1983, 1984b) an: Sie beschäftigen sich unter anderem mit der Analyse von Scheidungsraten. Wegen der gesetzlichen Regelungen ist eine gewisse Mindest–Wartezeit einzukalkulieren. Es ist auch wahrscheinlich, daß ein Großteil der Ehen erst mit dem Tod des einen Ehepartners aufgelöst wird. Ein Teil der Ehen wird also nie geschieden. Die Gompertz–Verteilung hat für diesen Fall die wünschenswerte Eigenschaft, daß S(t) für unendlich große Wartezeiten nicht Null sondern gleich $\exp(\lambda/\gamma)$ ist, wenn $\gamma$ negativ ist[1]. Da jedoch die Scheidungsrate nach der Heirat gering ist und erst nach einiger Zeit zunimmt, um dann wieder auf ein niedrigeres Niveau abzusinken, ist die Gompertz–Rate zur Modellierung dieses Prozesses nicht geeignet. Sie gestattet lediglich monoton zu– oder abnehmende Risiken (die Weibull–

1) Man beachte, daß GOMPERTZ (1825) ursprünglich von einer positiven Konstanten $\gamma$ ausgegangen ist.

Rate ebenfalls). In diesem Fall benötigt man also eine Verteilung mit wechselndem Risiko, z.B. die *Log-Normalverteilung* (vgl. Abbildung B.4). Da sie jedoch mathematisch nicht einfach zu handhaben ist, verwendet man häufig als Approximation die *log-logistische Verteilung* (vgl. Abbildung B.6). DIEKMANN und MITTER vergleichen diese Verteilung mit der von ihnen entwickelten Sichel-Funktion, für die gilt:

$$(5.54a) \quad f(t) = \exp\{-\lambda\gamma[\gamma - (t+\gamma)\exp(-t/\gamma)]\}\ \lambda t \exp(-t/\gamma)$$

$$(5.54b) \quad S(t) = \exp\{-\lambda\gamma[\gamma - (t+\gamma)\exp(-t/\gamma)]\}$$

$$(5.54c) \quad r(t) = \lambda t \exp(-t/\gamma)$$

Die Überlebensfunktion (5.54b) sinkt für $t \to \infty$ nicht auf den Wert Null, sondern entspricht $\exp(-\lambda\gamma^2) > 0$. Der zeitliche Verlauf der Rate sieht aus wie eine Sichel.

### 5.4.1.2 Flexible Modelllierung unterschiedlicher Zeitverläufe der Rate

Die *Exponentialverteilung* wird relativ häufig verwendet, weil sich dieses Modell besonders einfach anwenden und interpretieren läßt. Es hat aber den Nachteil, daß es von einer konstanten Rate ausgeht. Es zeigt sich sogar, daß die Schlußfolgerungen dieses Modells wesentlich von dieser Annahme abhängen. Anders ausgedrückt, die Exponentialverteilung ist gegenüber Verletzungen der Annahme nicht besonders robust.[1] Von daher stellt sich die Frage, ob man nicht die wesentlichen Eigenschaften des Modells beibehalten, es aber so flexibel machen kann, daß es verschiedene Veränderungen des Risikos im Zeitablauf berücksichtigt. Betrachten wir dazu den Teil des Modells (5.2), der die Zeitabhängigkeit des Prozesses beschreibt: die Basisrate $\lambda_0(t)$.

Eine Möglichkeit ist z.B., den gesamten Prozeß in mehrere Zeitintervalle zu zerstückeln, innerhalb derer die Rate konstant, zwischen denen aber Veränderungen zugelassen sind. Dieses Vorgehen läuft darauf hinaus, den

1) An Gleichung 5.7 erkennt man, daß exponentiell verteilte Wartezeiten eine bestimmte Varianz der Fehlerterme des Skalierungsmodells implizieren – nämlich $\sigma^2 = 1$. Die Restriktion des Parameters $\sigma$ auf den Wert 1 ist unter empirischen Gesichtspunkten nicht besonders plausibel.

empirischen Verlauf einer Rate durch eine Treppenfunktion anzunähern. Mit anderen Worten, der gesamte zeitabhängige Prozeß wird durch eine Aneinanderreihung zeitkonstanter Poisson-Prozesse approximiert. Je mehr Zeitintervalle man wählt, um so genauer wird diese Näherungslösung, um so mehr Parameter sind aber auch zu schätzen. Gegeben seien $p=1,\ldots,P$ Zeitintervalle $[\tau_p,\tau_{p+1})$ mit $t_1=0$. Eine zeitkonstante, *periodenspezifische Rate* läßt sich dann folgendermaßen definieren:

$$\lambda_0(t) = \lambda^p \quad \text{für } \tau_p \leq t < \tau_{p+1} \tag{5.55}$$

Die Basisüberlebensfunktion ist dann analog (5.40):

$$S_0(t) = \exp\left[\int_0^t \lambda_0(u)du\right] = \exp\left[-\sum_p^P \int_{\tau_p}^{\tau_{p+1}} \lambda^p du\right]$$

$$= \exp\left[-\sum_p^P \lambda^p(\tau_{p+1} - \tau_p)\right]$$

Mit der Intervalldauer $\Delta\tau_p=\tau_{p+1}-\tau_p$ läßt sich dieser Ausdruck schließlich vereinfachen zu:

$$S_0 = \prod_p^P \exp(-\lambda^p \Delta\tau_p) \tag{5.56}$$

Die Gesamtüberlebenswahrscheinlichkeit entspricht jetzt dem Produkt der periodenspezifischen Überlebensfunktionen.[1] Natürlich ist die Wahl der Zeitintervalle Forscherentscheidung und hat daher ein gewisses Maß an Beliebigkeit. Man wird dabei in vielen Fällen mehr nach technischen als nach inhaltlichen Gesichtspunkten vorgehen. Graphische Darstellungen der Rate können dabei weitere Entscheidungshilfen bieten (vgl. Abschnitt 4.5).

Statt durch eine Treppenfunktion kann man die Rate auch durch ein Polynom m-ten Grades approximieren:

---

1) Diese Feststellung läßt sich auch auf zeitabhängige, periodenspezifische Raten übertragen, wenn die zeitabhängige Rate lediglich eine Funktion der Intervalldauer $\Delta\tau_p$ (nicht der Prozeßzeit !) ist. Man könnte auch sagen, die Uhr des Prozesses muß mit jedem neuen Intervall wieder auf Null zurückgestellt werden.

(5.57a) $$\lambda_0(t) = \lambda_0 + \lambda_1 t + \lambda_2 t^2 + \ldots + \lambda_m t^m$$

Es ergibt sich eine verallgemeinerte Rayleigh-Verteilung (vgl. LAWLESS 1980: 252ff.) mit Überlebenswahrscheinlichkeit

(5.57b) $$S_0(t) = \exp\left[-\lambda_0 t - \frac{\lambda_1}{2}t^2 - \frac{\lambda_2}{3}t^3 - \ldots - \frac{\lambda_m}{m+1}t^{m+1}\right]$$

Der Nachteil dieses Verteilungsmodells ist, daß (5.57a) negative Werte ergeben kann (je nach Werten der $\lambda_m$), was für eine Rate ein unzulässiges Resultat ist. Diese Schwierigkeit kann dadurch umgangen werden, daß man auf der rechten Seite von (5.57a) eine Exponentialfunktion verwendet.

Eine andere Art der Flexibilisierung der Exponentialverteilung ist von WEIBULL (1951) vorgeschlagen worden. Hier ist die Rate eine Potenzfunktion der Zeit t:

(5.58) $$\lambda_0(t) = \lambda_\gamma(\lambda t)^{\gamma-1}$$

Man kann mit der *Weibull-Verteilung* die unterschiedlichsten Funktionsverläufe modellieren (vgl. Abbildung B.2), darunter konstante, linear, quadratisch (kubisch etc.) zu- und abnehmende Verläufe der Rate. Diese Vielfalt der Funktionsverläufe macht die Weibull-Verteilung zum am meisten verwendeten Analysemodell für Wartezeiten.[1] Da es gleichzeitig die Exponentialverteilung als Spezialfall enthält ($\gamma = 1$), kann man jederzeit die Annahme einer konstanten Rate testen.

Darüber hinaus ist die Weibull-Verteilung im Gegensatz zu allen anderen zeitabhängigen Verteilungsmodellen noch vergleichsweise einfach zu handhaben. Der Kehrwert des Parameters $\lambda$ ist genau der Zeitpunkt, den 36,8% der Stichprobe ohne Ereignis überleben, unabhängig davon, welchen Wert $\gamma$ bzw. welche Form die Verteilung hat. Das erkennt man leicht, wenn man in die Formel für S(t) der Weibull-Verteilung $1/\lambda$ einsetzt. Zur

1) Sind Wartezeiten Weibull-verteilt, dann ist ihr natürlicher Logarithmus extremwertverteilt (vgl. Abschnitt 2.3.3). Der Parameter $\sigma$ des Skalierungsmodells (5.8) ist jetzt nicht mehr auf den Wert 1 festgelegt, sondern entspricht $\gamma^{-1}$. Die Weibull-Verteilung ist also insofern eine Flexibilisierung der Exponentialverteilung, als sie die Varianz der Fehlerterme nicht festlegt – unter empirischen Gesichtspunkten eine sehr viel realistischere Perspektive.

Berechnung des Mittelwerts und der Varianz der Verteilung benötigt man die Gamma-Funktion, die z.B. bei ABRAMOWITZ und STEGUN (1965) tabelliert ist.

### 5.4.1.3 Diskriminierung unterschiedlicher Funktionsverläufe im Rahmen allgemeiner Verteilungsmodelle

Wenn die Exponentialverteilung ein Spezialfall der Weibull-Verteilung ist, ergibt sich allgemein die Frage, ob man nicht Verteilungsmodelle finden kann, die möglichst viele der in Anhang B genannten Verteilungen als Spezialfall enthalten. Auf diese Art und Weise könnte man, wenn keinerlei oder konkurrierende Informationen über den Verlauf der Rate vorliegen, verschiedene mögliche Modelle gegeneinander testen. Mit Hilfe des Weibull-Modells läßt sich ja nur prüfen, ob der Parameter $\gamma$ signifikant von Eins verschieden ist. Abweichungen können aber aus ganz unterschiedlichen Gründen zustandekommen und deshalb möchte man verschiedene Modelle und nicht nur die Exponentialverteilung prüfen.

Einen solchen Ansatz verfolgen KALBFLEISCH und PRENTICE (1980: 27ff., 63ff.) mit einer *verallgemeinerten Gamma-* bzw. einer *verallgemeinerten F-Verteilung*. Sie enthalten als Spezialfälle die Exponential-, die Gamma-, die Weibull- und die Log-Normalverteilung. Beide Verteilungsmodelle sind jedoch nur in wenigen Programmpaketen implementiert.

FLINN und HECKMAN (1982) benutzen eine ähnlich flexible Rate wie die Rayleigh-Verteilung (5.57). Statt eines Polynoms verwenden sie jedoch eine Box-Cox-Transformation:

$$\lambda_0(t) = \exp\left(\sum_{m=1}^{M} \lambda_m \frac{t^{\gamma_m} - 1}{\gamma_m}\right)$$

Der Vorteil dieser Transformation besteht darin, daß sich einige der bisher diskutierten Verteilungsmodelle als Spezialfälle ergeben, wenn lediglich $M = 1$ Summand verwendet wird:

– Gompertz–Verteilung ($M=1,\ \gamma_1=1$)

$$\lambda_0(t) = \exp\left[\lambda_1 \frac{t-1}{1}\right] = \exp(\lambda_1 t - \lambda_1) = \exp\lambda_1 \exp(\lambda_1 t)$$

– Weibull–Verteilung ($M=1,\ \gamma_1=0$)

$$\lambda_0(t) = \exp\left[\lambda_1 \frac{t^0-1}{0}\right] = \left\{\exp\left[\frac{t^0-1}{0}\right]\right\}^{\lambda_1} = [\exp(\ln t)]^{\lambda_1} = t^{\lambda_1}$$

$$\text{wegen } (t^{\gamma_1}-1)/\gamma_1 \approx \ln t \quad \text{für } |\gamma_1| \leq 10^{-6}$$

Die Parametrisierung weicht leicht von der in Anhang B gewählten Form ab – entscheidend ist jedoch, daß in einem Fall die Zeit als Exponentialfunktion (Gompertz) und im anderen Fall als Potenzfunktion (Weibull) vorkommt. Die Box–Cox–Transformation macht also die Basisrate $\lambda_0(t)$ hinreichend flexibel, um die verschiedensten Verläufe zu approximieren. Gleichzeitig ist es möglich, verschiedene Modelle gegeneinander zu testen.

### *5.4.2 Schätzung zeitabhängiger Modelle*

Nach dieser langen Vorrede möchte ich abschließend noch kurz demonstrieren, wie man ein zeitabhängiges Modell an Hand empirischer Daten überprüft. Im Prinzip stellen sich dabei keine neuen Probleme, man muß lediglich die übergangsspezifischen Raten und Überlebensfunktionen kennen und sie in die allgemeine Likelihood–Funktion (5.46) einsetzen. Für das Gompertz–Modell ergäbe sich z.B.:

$$L = \prod_j^J \prod_k^K \prod_i^N \exp\left[\frac{\lambda_{jk}}{\gamma_{jk}}[1-\exp(\gamma_{jk}t_i)]\right] [\lambda_{jk}\exp(\gamma_{jk}t_i)]^{\delta_{ijk}}$$

Wie in den vorhergehenden Abschnitten macht man auch bei zeitabhängigen Ratenmodellen einzelne Parameter von den Kovariaten abhängig, um die Heterogenität der Daten zu kontrollieren. Faßt man die Gompertz–Rate als proportionales Risiko–Modell auf

$$r_{jk}(t|\mathbf{x}_{ijk},\boldsymbol{\beta}_{jk}) = \lambda_{0jk}(t)\exp(\mathbf{x}_{ijk},\boldsymbol{\beta}_{jk}) \quad \text{mit } \lambda_{0jk}(t) = \exp(\gamma_{jk}t),$$

dann ist der Parameter $\lambda_{jk}$ des Gompertz-Modells (s.oben) eine log-lineare Funktion der Kovariaten. Entsprechend gilt für das Weibull-Modell[1]

$$r_{jk}(t|\mathbf{x}_{ijk},\boldsymbol{\beta}_{jk}) = \lambda_{0jk}(t)\exp(\mathbf{x}_{ijk}\boldsymbol{\beta}^{*}_{jk}) \quad \text{mit } \lambda_{0jk}(t) = \gamma_{jk}t^{\gamma_{jk}-1} \quad \text{und } \boldsymbol{\beta}^{*}_{jk} = \gamma_{jk}\boldsymbol{\beta}_{jk}$$

und das periodenspezifische Exponential-Modell

$$r_{jk}(t|\mathbf{x}^{p}_{ijk},\boldsymbol{\beta}^{p}_{jk}) = \lambda_{0jk}(t)\exp(\mathbf{x}^{p}_{ijk}\boldsymbol{\beta}^{p}_{jk}) \quad \text{mit } \lambda_{0jk}(t) = \lambda^{p}_{jk} \quad \text{für } \tau_{p} \leq t < \tau_{p+1}$$

Aus dem (hochgestellten) Index p erkennt man, daß in dem letzten Modell Kovariaten und Effekte je nach Periode variieren können.[2]

Alle drei Modelle sollen an Hand der MZ71-Daten illustriert werden. Die entsprechenden Schätzungen eines Gompertz-Modells 4, eines Weibull-Modells 5 und eines periodenspezifischen Exponentialmodells 6 enthält Tabelle 5.3 (PRG0501). Die Effekte aller Variablen, die bereits in Modell 3 untersucht wurden, bleiben insgesamt stabil. Auf Grund der Hypothesen ist zu erwarten, daß ein zeitabhängiges Modell insbesondere die Prognose sozialer Aufstiege verbessert. Der entsprechende Parameter $\gamma_{jk}$ des Gompertz- bzw. Weibull-Modells sollte ein positives Vorzeichen haben (vgl. Abbildung 3.1). Die Ergebnisse bestätigen beide Hypothesen. Allerdings zeigt sich auch bei den sozialen Abstiegen ein signifikant positiver Effekt der Tätigkeitsdauer bei zugegebenermaßen geringerer Verbesserung des Modellfits relativ zu Modell 3. Dies mag damit zusammenhängen, daß soziale Abstiege mit den vorliegenden Variablen nicht besonders gut prognostiziert werden können und die zeitabhängigen Parameter daher einen Teil der unbeobachteten Heterogenität messen (vgl. Abschnitt 2.5.1).

Angesichts einer erheblich größeren Parameterzahl erzielt das periodenspezifische Exponentialmodell 6 den besten Modellfit. Es geht davon aus, daß der Effekt aller vier Kovariaten über alle Perioden konstant ist, während die Regressionskonstante für jede der sechs zweijährigen Intervalle

---

1) Grundlage ist die folgende Umformung von (5.58): $r(t)=\lambda^{\gamma}\gamma t^{\gamma-1}$. Wegen $\lambda^{\gamma}$ ist der Parameter $\lambda$ der Weibull-Rate nicht zwangsläufig (wie bei der Gompertz-Rate) eine einfache log-lineare Funktion der Kovariaten. Bei Verwendung von Computer-Programmen ist sehr genau darauf zu achten, wie das Programm parametrisiert. Ich habe eine Form gewählt, bei der nur $\lambda$ eine log-lineare Funktion der Kovariaten ist. Diese Parametrisierung ist mit den Programmen SAS und TDA kompatibel (vgl. Anhang D).

2) Auf diese Art und Weise kann man übrigens auch *zeitabhängige Kovariaten* berücksichtigen (vgl. Abschnitt 6.1).

*Tabelle 5.3: Modelle mit zeitabhängiger Rate (MZ71 – Daten)*

| Variable | | Abstiege Modell 3 | Modell 4 | Modell 5 | Modell 6 | Aufstiege Modell 3 | Modell 4 | Modell 5 | Modell 6 |
|---|---|---|---|---|---|---|---|---|---|
| Konstante | a) | −2,7030* | −3,0256* | −2,6487* | | −5,7443* | −6,7003* | −4,5600* | |
| | b) | 0,2876 | 0,2997 | 0,2278 | | 0,1236 | 0,1347 | 0,0850 | |
| | c) | 0,0670 | 0,0485 | 0,0707 | | 0,0032 | 0,0012 | 0,0105 | |
| Status des | | −0,0011 | −0,0011 | −0,0008* | −0,0010 | 0,0029* | 0,0030* | 0,0019* | 0,0031* |
| Vaters | | 0,0006 | 0,0006 | 0,0004 | 0,0006 | 0,0002 | 0,0002 | 0,0002 | 0,0002 |
| | | 0,9989 | 0,9989 | 0,9992 | 0,9990 | 1,0029 | 1,0030 | 1,0019 | 1,0031 |
| Ausbildung des | | −0,1789* | −0,1728* | −0,1343* | −0,1685* | 0,2827* | 0,3105* | 0,1934* | 0,3200* |
| Befragten | | 0,0299 | 0,0302 | 0,0244 | 0,0303 | 0,0117 | 0,0117 | 0,0078 | 0,0117 |
| | | 0,8362 | 0,8413 | 0,8743 | 0,8449 | 1,3267 | 1,3641 | 1,2134 | 1,3771 |
| Status des | | 0,0021* | 0,0019* | 0,0014* | 0,0017* | −0,0118* | −0,0128* | −0,0080* | −0,0132* |
| Befragten | | 0,0009 | 0,0009 | 0,0007 | 0,0009 | 0,0005 | 0,0005 | 0,0003 | 0,0005 |
| | | 1,0021 | 1,0019 | 1,0014 | 1,0017 | 0,9883 | 0,9873 | 0,9920 | 0,9869 |
| Beschäftigten – | | 0,0010 | 0,0011 | 0,0009 | 0,0011 | 0,0012* | 0,0014* | 0,0009* | 0,0015* |
| wachstum | | 0,0012 | 0,0012 | 0,0010 | 0,0012 | 0,0006 | 0,0006 | 0,0004 | 0,0006 |
| Ausgangsbranche | | 1,0010 | 1,0011 | 1,0009 | 1,0011 | 1,0012 | 1,0014 | 1,0009 | 1,0015 |
| Tätigkeits – | | | 0,0555* | 0,2404* | | | 0,1382* | 0,4713* | |
| dauer | | | 0,0130 | 0,0355 | | | 0,0070 | 0,0182 | |
| | | | 1,0571 | 1,2718 | | | 1,1482 | 1,6021 | |
| Periode 1 | | | | | −3,4933* | | | | −7,4795* |
| | | | | | 0,3104 | | | | 0,1473 |
| | | | | | 0,0304 | | | | 0,0006 |
| Periode 2 | | | | | −2,6597* | | | | −6,2974* |
| | | | | | 0,2999 | | | | 0,1335 |
| | | | | | 0,0700 | | | | 0,0018 |
| Periode 3 | | | | | −2,8685* | | | | −6,0244* |
| | | | | | 0,3021 | | | | 0,1317 |
| | | | | | 0,0568 | | | | 0,0024 |
| Periode 4 | | | | | −2,8020* | | | | −5,7028* |
| | | | | | 0,3013 | | | | 0,1287 |
| | | | | | 0,0607 | | | | 0,0033 |
| Periode 5 | | | | | −2,3354* | | | | −5,4897* |
| | | | | | 0,2981 | | | | 0,1275 |
| | | | | | 0,0968 | | | | 0,0041 |
| Periode 6 | | | | | −2,1035* | | | | −5,1340* |
| | | | | | 0,3002 | | | | 0,1457 |
| | | | | | 0,1220 | | | | 0,0059 |
| N | | 7587 | 7587 | 7587 | 7587 | 7587 | 7587 | 7587 | 7587 |
| D | | 690 | 690 | 690 | 690 | 2448 | 2448 | 2448 | 2448 |
| in % | | 9,1% | 9,1% | 9,1% | 9,1% | 32,3% | 32,3% | 32,3% | 32,3% |
| ln L | | −3727,18 | −3718,10 | −3706,16 | −3679,09 | −9725,58 | −9532,63 | −9445,43 | −9282,97 |
| Pseudo – $R^2$ | | 0,9% | 1,1% | 1,4% | 2,1% | 5,0% | 6,9% | 7,7% | 9,3% |
| L – Ratio | | 63,97 | 82,14 | 106,01 | 160,16 | 1022,65 | 1408,53 | 1582,95 | 1907,85 |
| df | | 4 | 5 | 5 | 9 | 4 | 5 | 5 | 9 |

a) Parameter (* sign. 5% – Niveau), b) Standardfehler, c) Antilogarithmus

verschieden, aber zeitkonstant sein soll. Die sechs Konstanten (Periode 1−6) zeigen sowohl für Aufstiege als auch für Abstiege (mit Ausnahme der 4. Periode) eine zunehmende Rate. Im Rahmen dieses Modelltyps kann man auch den Effekt des Krisenjahres 1967 testen. Wenn die untersuchten Berufsanfänger in den Jahren 1961−63 in den Arbeitsmarkt eingetreten sind und im Jahr 1967 vermehrt soziale Abstiege auftreten sollten, dann sollte 4−6 Jahre nach Eintritt in das Berufsleben, die Abstiegsrate signifikant höher sein als in den beiden Perioden davor (0−3 Jahre) und danach (7−11 Jahre). Die geschätzten Regressionskonstanten des entsprechenden 3−Perioden−Modells (−2,9916, −2,7866, −2,2837) unterstützen diese These nicht.

## 5.5 *Partiell parametrische Regressionsmodelle − Partial Likelihood*

In diesem Abschnitt möchte ich mich eingehender mit dem von COX (1972) vorgeschlagenen Regressionsmodell beschäftigen. Der besondere Vorzug dieses Modells ist, daß es keine spezielle Aussage über die Art der Zeitabhängigkeit macht. Dennoch werden Veränderungen im Zeitablauf kontrolliert: Sie werden quasi aus den Daten herausgerechnet. Dieses Modell ist deshalb attraktiv,

- weil man manchmal keine Informationen über die Art der Veränderung hat,
- weil der empirische Verlauf der Rate so unregelmäßig sein kann, daß er durch eine mathematische Funktion nur sehr schlecht angenähert wird oder
- weil man hauptsächlich an dem Einfluß der Kovariaten interessiert ist und Veränderungen im Zeitablauf in der Tat nur kontrollieren möchte.

Da das Modell jedoch nur teilweise spezifiziert ist, sind spezielle Methoden notwendig, um es an Hand empirischer Daten zu überprüfen. In Abschnitt 5.5.1 werde ich das prinzipielle Vorgehen dieser sogenannten PL−Schätzungen wiederum an Hand eines empirischen Beispiels erläutern. Dabei zeigt sich, daß der in Abschnitt 4.4.2 besprochene verallgemeinerte Savage−Test nur ein Spezialfall dieses Ansatzes ist.

Zur Berechnung der PL−Schätzer verwendet man die Rangordnung aller Wartezeiten. Wie schon bei der Besprechung der o.g. Rangtests deut-

lich wurde, ist dabei das Auftreten von ties (Ereignisse zum gleichen Zeitpunkt) ein besonderes Problem. Bei einer geringen Anzahl von ties ist eine geringfügige Abänderung der PL – Schätzer möglich, die ich ebenfalls in Abschnitt 5.5.1 besprechen werde. Treten jedoch wie bei den MZ71 – Daten sehr viele Ereignisse zum gleichen Zeitpunkt auf, dann ist es sinnvoller, gleich ein *PL – Modell für diskrete Wartezeiten* zu verwenden. Dessen Behandlung würde aber den Rahmen dieses Buches sprengen (vgl. jedoch LAWLESS 1982: 372ff. und KALBFLEISCH/PRENTICE 1980: 98ff.). Bei diesem Datensatz sind also die Anwendungsvoraussetzungen für ein kontinuierliches PL – Modell nicht gegeben. Um dennoch einen Eindruck von der Praktikabilität dieses Ansatzes zu vermitteln, werde ich in Abschnitt 5.5.2.1 das Rückfallrisiko der 64 Strafentlassenen (KFN – Daten) untersuchen.

COX erfaßt Veränderungen im Zeitablauf durch eine nicht näher spezifizierte Basisrate $\lambda_0(t)$. Um diese Unbekannte aus den Daten zu eliminieren, stützen sich PL – Schätzungen auf die Annahme eines proportionalen Risikos. Da diese Annahme grundlegende Voraussetzung der weiteren Berechnungen ist, ist es hier schon zu einem frühen Zeitpunkt der Datenanalyse notwendig, die Existenz proportionaler Risiken zu prüfen. Ich werde hierzu in Abschnitt 5.5.2.2 einen einfachen graphischen und einen mehr formellen inferenzstatistischen Test vorstellen. Ist die Annahme proportionaler Risiken nicht gegeben, dann ist das Modell von COX nicht unbedingt hinfällig, vorausgesetzt, man kann die Stichprobe in mehrere Teilgruppen zerlegen, innerhalb derer die Proportionalitätsannahme gegeben ist.

### *5.5.1 Partial – Likelihood – Schätzungen*

Grundlage der folgenden Berechnungen ist das Modell proportionaler Risiken (5.2). Der Einfachheit halber betrachte ich zunächst einmal singuläre Ereignisse und modelliere daher die Rate r(t). Um eine ML – Schätzung durchführen zu können, muß man, wie in Abschnitt 5.2 beschrieben, die Wahrscheinlichkeitsdichte jeder Beobachtung unter den Modellannahmen angeben können. Das ist in diesem Fall solange nicht möglich, wie die Basisrate $\lambda_0(t)$ unbekannt bleibt. COX hat daher den Vorschlag gemacht, den Beitrag jedes Ereignisses zur Gesamtlikelihood durch folgenden Ausdruck zu messen:

(5.60a) $$PL_\iota = \frac{\exp(\mathbf{x}_\iota\beta)}{\sum_{l \in R(t_\iota)} \exp(\mathbf{x}_l\beta)}$$

$t_\iota$ sei dabei der Zeitpunkt, zu dem das $\iota$-te Ereignis stattfindet. $R(t_\iota)$ sei die *Risikomenge* zum gleichen Zeitpunkt, also die Anzahl der Untersuchungseinheiten, die noch ein Ereignis haben können. Der Parametervektor $\beta$ und der Zeilenvektor $\mathbf{x}_\iota$ der Kovariaten sind definiert wie bisher. Sollten zum gleichen Zeitpunkt mehrere Ereignisse auftreten, die aber relativ zum Stichprobenumfang nicht ins Gewicht fallen, dann ist folgende Näherungsformel sinnvoll. Dabei sei $D(t_\iota)$ die Menge der Personen, die zum Zeitpunkt $t_\iota$ ausscheiden. Insgesamt seien es $d_\iota$ Personen.

(5.60b) $$PL_\iota = \frac{\exp(\mathbf{z}_\iota\beta)}{\left(\sum_{l \in R(t_\iota)} \exp(\mathbf{x}_l\beta)\right)^{d_\iota}} \quad \text{mit } \mathbf{z}_\iota = \sum_{l \in D(t_\iota)} \mathbf{x}_l$$

COX hat weiterhin vorgeschlagen, das Produkt dieser Einzelwahrscheinlichkeiten wie eine Likelihood-Funktion im üblichen Sinne zu behandeln, obwohl in (5.60) keine Aussagen über zensierte Beobachtungen gemacht werden. Seine erste Begründung dieses Vorgehens (1972) beruhte auf folgender Überlegung: Wenn die Basisrate unbekannt ist, erhält man keine zusätzlichen Informationen über $\beta$ durch Kenntnis des Prozesses zwischen den Ereignissen. Man könnte die Basisrate zwischen den Ereignissen einfach Null setzen und nichts würde sich ändern. Folglich sei es gerechtfertigt, sich bei PL-Schätzungen ausschließlich auf Ereignisse zu konzentrieren. Dabei entspricht der Ausdruck (5.60) der bedingten Wahrscheinlichkeit, daß von $R(t_\iota)$ Personen genau $d_\iota$ Personen ausscheiden. In einem weiteren Aufsatz (1975) hat COX dann diesen Ansatz unter formalen Gesichtspunkten untersucht und gezeigt, daß es sich bei dem Produkt der bedingten Wahrscheinlichkeiten (5.60) um eine *partielle Likelihood* handelt.

Ich möchte in der gleichen Reihenfolge vorgehen und den Vorschlag zunächst unhinterfragt übernehmen. In Abschnitt 5.5.1.1 werde ich die konkrete Berechnung von PL-Schätzern an Hand eines früheren Beispiels demonstrieren. Im folgenden Abschnitt 5.5.1.2 werde ich dann zeigen, daß es sich bei dem o.g. Produkt um einen Teil der gesamten Likelihood-Funktion handelt, wie sie aus vorhergehenden Abschnitten bekannt ist. Es zeigt sich, daß die Überlebenswahrscheinlichkeit vernachlässigt wird, weil ihre Berechnung die Kenntnis der Basisrate $\lambda_0(t)$ voraussetzt (vgl. Gleichung

5.4). In Abschnitt 5.5.1.3 werde ich kurz auf das Problem der ties eingehen und die Näherungsformel (5.60b) motivieren. Schließlich stellt sich in Abschnitt 5.5.1.4 die Frage, ob man, nachdem die PL – Schätzer der Regressionskoeffizienten bekannt sind, nicht doch noch Angaben über die Überlebensfunktion S(t) machen kann.

### 5.5.1.1 Berechnung der PL – Schätzer an Hand eines Beispiels

Zunächst einmal möchte ich zeigen, daß es sich bei dem Ausdruck (5.60a) um eine bedingte Wahrscheinlichkeit handelt und daß durch diese Betrachtung die unspezifizierte Basisrate $\lambda_0(t)$ eliminiert werden kann, vorausgesetzt, es handelt sich um ein Ratenmodell mit proportionalen Risiken. Angenommen die Risikomenge besteht nur noch aus zwei Personen A und B und zum Zeitpunkt $t_i$ scheidet Person A aus. Gesucht ist also die bedingte Wahrscheinlichkeit, daß Person A ausscheidet, vorausgesetzt es findet überhaupt ein Ereignis statt.

Die Wahrscheinlichkeit, daß A im Intervall $[t,t+dt)$ ein Ereignis aufweist, B aber nicht, entspricht dem Produkt der Dichte für A und der Überlebenswahrscheinlichkeit für B:

$$\text{(5.61a)} \qquad P(\text{Ereignis A}) = f_a(t_i)dt\ S_b(t_i)$$

Zur besseren Unterscheidung werden die beiden Funktionen mit den Indizes a und b gekennzeichnet. Wäre B statt A ausgeschieden, würde der Ausdruck dementsprechend lauten:

$$\text{(5.61b)} \qquad P(\text{Ereignis B}) = f_b(t_i)dt\ S_a(t_i)$$

Die Wahrscheinlichkeit, daß überhaupt ein Ereignis zum Zeitpunkt $t_i$ auftritt, entspricht der Summe der Einzelwahrscheinlichkeiten (5.61a) und (5.61b):

$$\text{(5.62)} \qquad P(\text{Ereignis A oder B}) = f_a(t_i)dt\ S_b(t_i) + f_b(t_i)dt\ S_a(t_i)$$

Die gesuchte bedingte Wahrscheinlichkeit entspricht dem Verhältnis von (5.61a) und (5.62):

(5.63a) P(Ereignis A|Ereignis A oder B)

$$= \frac{f_a(t_i)dt\ S_b(t_i)}{f_a(t_i)dt\ S_b(t_i) + f_b(t_i)dt\ S_a(t_i)}$$

Unter Verwendung von (2.14) und durch Einsetzen des Regressionsmodells (5.59) erhält man (dt gekürzt):

(5.63b) P(Ereignis A|Ereignis A oder B)

$$= \frac{r_a(t_i)S_a(t_i)S_b(t_i)}{r_a(t_i)S_a(t_i)S_b(t_i) + r_b(t_i)S_b(t_i)S_a(t_i)}$$

$$= \frac{r_a(t_i)}{r_a(t_i) + r_b(t_i)}$$

$$= \frac{\lambda_0(t_i)\exp(\mathbf{x}_a\boldsymbol{\beta})}{\lambda_0(t_i)\exp(\mathbf{x}_a\boldsymbol{\beta}) + \lambda_0(t_i)\exp(\mathbf{x}_b\boldsymbol{\beta})}$$

$$= \frac{\exp(\mathbf{x}_a\boldsymbol{\beta})}{\exp(\mathbf{x}_a\boldsymbol{\beta}) + \exp(\mathbf{x}_b\boldsymbol{\beta})}$$

$$= \frac{\exp(\mathbf{x}_a\boldsymbol{\beta})}{\sum_{l=a,b}\exp(\mathbf{x}_l\boldsymbol{\beta})}$$

Gleichung (5.63b) entspricht dem Ausdruck (5.60a). In diesem Fall besteht die Risikomenge nur noch aus den Personen A und B. Durch Betrachtung einer bedingten Wahrscheinlichkeit kann man also die unspezifizierte Basisrate kürzen[1]. Man beachte, daß das bei einem additiven Modell nicht möglich gewesen wäre. Anders ausgedrückt, um die *unbekannte Basisrate eliminieren* zu können, muß man ein *Modell mit proportionalem Risiko* betrachten.

Zur konkreten Berechnung einer PL−Schätzung verwende ich noch einmal das Beispiel aus Tabelle 4.4: Mich interessiert die Frage, ob Personen mit sozialen Problemen ein sehr viel höheres Rückfallrisiko haben als

1) Voraussetzung dafür ist, daß die Basisrate für alle Untersuchungseinheiten gleich ist. Man beachte außerdem, daß das gleiche Argument auf die Regressionskonstante $\beta_0$ zutrifft. PL−Modelle enthalten daher kein konstantes Glied bzw. dieser Parameter ist von vorneherein auf 0 festgelegt.

der Rest der Stichprobe. Zu diesem Zweck definiere ich eine Dummy-Variable $x_{i1}$, die nur dann den Wert 1 hat, wenn eine Person Probleme im sozialen Umfeld hat (sonst 0). Das Regressionsmodell lautet dementsprechend:

$$r(t) = \lambda_0(t) \exp(\beta_1 x_{i1}) \tag{5.64}$$

mit $r(t|x_{i1}=1) = \lambda_0(t) \exp \beta_1$ (Gruppe 1: mit Problemen)
und $r(t|x_{i1}=0) = \lambda_0(t)$ (Gruppe 2: ohne Probleme)

Die Basisrate mißt jetzt das Rückfallrisiko der Restgruppe, während der Parameter $\beta_1$ angibt, um wieviel das Rückfallrisiko der Personen mit sozialen Problemen höher ist. Die (partielle) Likelihood-Funktion lautet:

$$PL = \prod_{\iota}^{I} \frac{\exp(\beta_1 z_{\iota 1})}{\left( \sum_{l \in R(t_\iota)} \exp(\beta_1 x_{l1}) \right)^{d_\iota}}$$

Man beachte, daß zensierte Beobachtungen ignoriert werden und daher nur die Terme für die verbleibenden D Ereignisse betrachtet werden müssen. Die Multiplikation erfolgt daher nur für die $\iota = 1,\ldots,I$ Zeitpunkte $t_\iota$, zu denen Ereignisse auftreten.[1] Treten keine ties auf, entspricht die Anzahl I der Ereigniszeitpunkte der Anzahl D der Ereignisse. Im allgemeinen Fall ist jedoch von ties auszugehen ($I \leq D$) und dementsprechend wurde die Näherungsformel (5.60b) verwendet.

Man maximiert wiederum die logarithmierte (partielle) Likelihood-Funktion:

1) Da man gleichzeitig die Risikomenge zum Zeitpunkt $t_\iota$ kennen muß, werden die Daten sinnvollerweise in aufsteigender Reihenfolge sortiert (vgl. z.B. Tabelle 4.3). Treten viele Ereignisse zum gleichen Zeitpunkt auf, dann ergibt sich wie in dem Beispiel eine zusammengefaßte Tabelle (vgl. Tabelle 4.4). Alle EDV-Programme für das Regressionsmodell von COX sortieren daher die Daten, was möglicherweise die Fallzahl beschränkt (vgl. die Diskussion des Kaplan-Meier-Schätzers in Abschnitt 4.3.1).

$$\text{(5.65a)} \qquad \ln PL = \beta_1 \sum_{i}^{I} z_{i1} - \sum_{i}^{I} \left\{ d_i \ln \left( \sum_{l \in R(t_i)} \exp(\beta_1 x_{l1}) \right) \right\}$$

$$\text{mit } \sum_{i}^{I} z_{i1} = D_1 \text{ und } \sum_{l \in R(t_i)} \exp(\beta_1 x_{l1}) = n_{i1} \exp \beta_1 + n_{i2}$$

Wie man sich leicht überlegen kann, entspricht der erste Summenterm der Gesamtzahl $D_1$ aller Ereignisse in der Gruppe 1 (mit Problemen), während der letzte Summenterm von dem Parameter $\beta_1$ und den gruppenspezifischen Risikomengen $n_{i1}$ und $n_{i2}$ zum Zeitpunkt $t_i$ abhängt. Auf diese Art und Weise läßt sich die (logarithmierte) Likelihood-Funktion weiter vereinfachen:

$$\text{(5.65b)} \qquad \ln PL = \beta_1 D_1 - \sum_{i}^{I} [d_i \ln (n_{i1} \exp \beta_1 + n_{i2})]$$

Die ersten und die (mit $-1$ malgenommenen) zweiten Ableitungen nach $\beta_1$ lauten dann:

$$\text{(5.66)} \qquad \Theta(\beta_1) = \frac{\partial \ln PL}{\partial \beta_1} = D_1 - \sum_{i}^{I} \frac{d_i n_{i1} \exp \beta_1}{n_{i1} \exp \beta_1 + n_{i2}}$$

$$\text{(5.67)} \qquad I(\beta_1) = \frac{-\partial^2 \ln PL}{\partial \beta_1 \, \partial \beta_1} = \sum_{i}^{I} \frac{d_i n_{i1} n_{i2} \exp \beta_1}{(n_{i1} \exp \beta_1 + n_{i2})^2}$$

Durch Nullsetzen von (5.66) erhält man die *gesuchten PL-Schätzer*. In diesem Fall muß also der folgende Ausdruck gleich der Anzahl aller Ereignisse in der Gruppe 1 (mit Problemen) sein (Daten s. Tabelle 4.4):

$$D_1 = \sum_{i}^{I} \frac{d_i n_{i1} \exp \beta_1}{n_{i1} \exp \beta_1 + n_{i2}} = \sum_{i}^{I} \frac{d_i n_{i1} \hat{a}_1}{n_{i1} \hat{a}_1 + n_{i2}}$$

$$15 = \frac{4 \cdot 18\hat{a}_1}{18\hat{a} + 46} + \frac{5 \cdot 14\hat{a}_1}{14\hat{a}_1 + 46} + \ldots + \frac{1 \cdot 1\hat{a}_1}{1\hat{a}_1 + 25} + \frac{1 \cdot 1\hat{a}_1}{1\hat{a}_1 + 22}$$

Diese Gleichung kann nicht so einfach nach $\hat{\beta}_1 = \ln \hat{a}_1$ aufgelöst werden. Durch Probieren erhält man jedoch einen Wert von 1,7316, der die Gleichung erfüllt. Dies ist der gesuchte PL–Schätzer (vgl. auch Tabelle 5.5 (Modell 1) sowie PRG0502). Der positive Wert für $\hat{\beta}_1$ zeigt, daß das Abstiegsrisiko für Personen ohne abgeschlossene Berufsausbildung höher ist.

Man benötigt also schon bei einem relativ einfachen Analyseproblem ein iteratives Computerprogramm. Man erinnere sich jedoch an den *Score–Test* aus Abschnitt 5.2.4. Sein Vorteil ist, daß man nicht erst die gesuchten Schätzwerte berechnen muß, sondern gleich die Parameter der Nullhypothese einsetzen kann. Testen wir also einmal die Nullhypothese, daß zwischen beiden Gruppen kein Unterschied besteht. In diesem Fall wäre $\tilde{\beta}_1 = 0$ und die Score–Funktion (5.66) reduziert sich auf:

$$(5.68) \quad \Theta(\tilde{\beta}_1 = 0) = D_1 - \sum_{t}^{I} \frac{d_t n_{t1}}{n_{t1} + n_{t2}} \quad \text{mit } I(\tilde{\beta}_1 = 0) = \sum_{t}^{\beta} \frac{d_t n_{t1} n_{t2}}{(n_{t1} + n_{t2})^2}$$

Unter Verwendung von (5.33) könnte man jetzt diese Nullhypothese testen. Man beachte jedoch, daß sich Gleichung (5.68) auch folgendermaßen schreiben läßt:

$$(5.69) \quad \Theta(\tilde{\beta}_1 = 0) = \sum_{t}^{I} \left( d_{t1} - \frac{d_t n_{t1}}{n_{t.}} \right)$$

Dieser Ausdruck ist nichts anderes als die Testgröße des *verallgemeinerten Savage–Testes*, wie ich sie in (4.24a) definiert habe.

Schließlich sei noch einmal darauf hingewiesen, daß die Likelihood–Funktion (5.64) natürlich von einer geringen Anzahl ties ausgeht. Für diskrete Wartezeiten mit vielen ties wäre entsprechend ein zeitdiskretes Regressionsmodell zu verwenden. Dabei ergibt sich eine andere Varianzformel (vgl. z.B. LAWLESS 1982: 380ff.), die der o.g. Formel (4.25a) des Savage–Tests entspricht.

### 5.5.1.2 Partial Likelihood

In diesem Abschnitt möchte ich zeigen, daß es sich bei der Likelihood–Funktion (5.64) nur um einen Teil der gesamten Likelihood handelt, wie sie in allen vorhergehenden Abschnitten verwendet wurde. Der Einfachheit halber verwende ich die Likelihood–Funktion für singuläre Ereignisse und setze in (5.18) das COX'sche Regressionsmodell ein. Für die Überlebenswahrscheinlichkeit verwende ich Gleichung (5.4b). Nach einigen Umformungen ergibt sich:

$$(5.70a) \qquad L = \prod_{i}^{N} S_0(t)^{\exp(\mathbf{x}_i\boldsymbol{\beta})} [\lambda_0(t) \exp(\mathbf{x}_i\boldsymbol{\beta})]^{\delta_i}$$

Für die folgende Ableitung will ich davon ausgehen, daß keine ties auftreten. Es finden also zu den Zeitpunkten $t_i$ $(i=1,\ldots,D)$ insgesamt D Ereignisse statt. C Beobachtungen seien zensiert $(N=C+D)$, wobei die entsprechenden Zensierungen niemals gleichzeitig mit einem anderen Ereignis auftreten, sondern ggfs. etwas später stattfinden (eindeutige Rangordnung, s. Abschnitt 4.3ff.). Folgende Umformung von (5.70a) verändert den Ausdruck nicht, macht jedoch die Bestandteile der Gesamt–Likelihood deutlich:

$$(5.70b)\ L = \prod_{i}^{N} \left( \frac{\exp(\mathbf{x}_i\boldsymbol{\beta})}{\sum_{l \in R(t_i)} \exp(\mathbf{x}_l\boldsymbol{\beta})} \right)^{\delta_i} \cdot \left( \lambda_0(t) \sum_{l \in R(t_i)} \exp(\mathbf{x}_l\boldsymbol{\beta}) \right)^{\delta_i} \cdot \left( S_0(t)^{\exp(\mathbf{x}_i\boldsymbol{\beta})} \right)$$

(5.70b) besteht danach aus drei Termen, von denen der erste der obigen partiellen Likelihood (5.60a) entspricht. Für den Rest dieser Likelihood–Funktion kann man zeigen (EFRON 1977), daß er

1. die Wahrscheinlichkeit beschreibt, daß genau zu den Zeitpunkten $t_i$ ein Ereignis stattfindet und dazwischen keines, und daß
2. keine weiteren Ereignisse nach $t_D$ stattfinden.

Mit anderen Worten, *PL–Schätzungen* berücksichtigen lediglich die Wahrscheinlichkeit, daß von einer gegebenen Risikomenge ein bestimmtes Individuum ein Ereignis hat, ignorieren aber die Zeitpunkte der einzelnen Ereignisse sowie alle zensierten Beobachtungen. Man beachte allerdings, daß

zensierte Beobachtungen nicht völlig vernachlässigt werden, denn bis zum Zeitpunkt ihrer Zensierung sind sie Teil der Risikomenge.

Für die Berechnung von PL−Schätzern muß man daher nur zu jedem Zeitpunkt die Risikomenge kennen, aus der dann sukzessive einzelne Personen ausscheiden. Diese Information erhält man am einfachsten durch eine Sortierung aller Wartezeiten in aufsteigender Reihenfolge (vgl. Tabelle 4.3). Anders ausgedrückt, PL−Schätzungen verwenden lediglich die Rangordnung der Daten.

Es zeigt sich, daß PL−Schätzungen unter einigermaßen allgemeinen Bedingungen die gleichen asymptotischen Eigenschaften haben wie ML−Schätzungen, obwohl sie einen Teil der gesamten Likelihood ignorieren (COX 1975, EFRON 1977, OAKES 1977). Alle Ergebnisse der ML−Theorie, insbesondere die o.g. Signifikanztests, sind also direkt übertragbar. Unerforscht ist auch hier das Verhalten der Schätzungen in kleinen Stichproben. Einige Simulationsexperimente mit Verlaufsdaten finden sich dazu bei TUMA (1982b).

Der PL−Ansatz kann auch auf multiple Ereignisse verallgemeinert werden, indem man entweder die jeweils nicht interessierenden Ereignisse zensiert (vgl. Abschnitt 5.2.6) oder bei der Risikomenge berücksichtigt, in welchem Zustand die Individuen sich jeweils befinden und wohin ein Wechsel stattfindet. Unter der Voraussetzung, daß die verschiedenen Risiken voneinander unabhängig sind, kann man eine *(partielle) Likelihood−Funktion für multiple Ereignisse* wie folgt schreiben (vgl. HOLT 1978):

$$(5.71) \qquad PL = \prod_{j}^{J} \prod_{k}^{K} \prod_{\iota}^{I} \left( \frac{\exp(\mathbf{x}_{ijk}\boldsymbol{\beta}_{jk})}{\sum_{l \in R_j(t_\iota)} \exp(\mathbf{x}_{ljk}\boldsymbol{\beta}_{jk})} \right)^{\delta_{ijk}}$$

$$\text{mit } \delta_{ijk} = \begin{cases} 1 & \text{(Wechsel von j nach k)} \\ 0 & \text{(Zensierung)} \end{cases}$$

Man beachte, daß die Risikomenge jetzt für jeden Herkunftszustand getrennt berechnet werden muß, daher die Bezeichnung $R_j(t_\iota)$. Es ist damit nicht mehr nötig, eine Rangordnung für die gesamte Stichprobe zu berechnen. Auf diese Weise vereinfacht sich die praktische Berechnung der Likelihood, obwohl die Formel (5.71) komplizierter aussieht.

### 5.5.1.3 Ereignisse zum gleichen Zeitpunkt (ties)

In diesem Abschnitt möchte ich erläutern, warum die Existenz von *ties* die Berechnungen erschwert und wie die Näherungsformel (5.60b) zustandekommt. Man erinnere sich dazu noch einmal an die obige Ableitung der bedingten Wahrscheinlichkeit (vgl. Gleichung 5.63). Angenommen die Risikomenge besteht aus 10 Personen und zum Zeitpunkt $t_i$ scheiden 5 Personen aus. Wollte man jetzt eine bedingte Wahrscheinlichkeit berechnen, dann müßte man berücksichtigen, wie viele Möglichkeiten es gibt, daß von 10 Personen 5 ein Ereignis haben. Dies ist ein Problem der Kombinatorik und wird ebenfalls von COX (1972) diskutiert.

PETO (1972) wählt einen etwas anderen Ansatz: Obwohl die 5 Ereignisse alle zum gleichen Zeitpunkt beobachtet werden, finden sie, seiner Annahme nach, tatsächlich nacheinander statt. Die bedingte Wahrscheinlichkeit ergibt sich also durch das Produkt der einzelnen (bedingten) Wahrscheinlichkeiten, daß erst Person 1, dann Person 2 usw. ein Ereignis haben. Auch hier muß man kombinatorisch die verschiedenen möglichen Reihenfolgen berücksichtigen.

Egal welchen der beiden Ansätze man wählt, schon bei mehr als zwei Ereignissen sind die folgenden Berechnungen sehr kompliziert und es ist ein diskretes PL−Modell vorzuziehen. Wenn der relative Anteil der ties gering ist, unterscheiden sich beide Ansätze nicht wesentlich von dem folgenden Ausdruck:

$$(5.72) \qquad PL_i = \frac{\exp(\mathbf{z}_i\boldsymbol{\beta})}{\varkappa \left( \sum_{l \in R(t_i)} \exp(\mathbf{x}_l\boldsymbol{\beta}) \right)^{d_i}} \qquad \text{mit } \varkappa = \binom{n_i}{d_i} n_i^{-d_i}$$

Dabei mißt $n_i$ die Anzahl der Elemente der Risikomenge $R(t_i)$ zum Zeitpunkt $t_i$. Der gesamte Ausdruck entspricht genau der obigen Näherungsformel (5.60b), wenn man berücksichtigt, daß der Faktor $\varkappa$ keinen Einfluß auf das Maximierungsproblem, also die Berechnung der PL−Schätzer hat. Das gleiche Ergebnis findet sich bei BRESLOW (1974), allerdings mit einer etwas anderen Begründung. BRESLOW's Näherungsformel wird in den meisten Programmen verwendet. In SAS kann man die Verwendung von ties durch eine besondere Option steuern. Neben der exakten Methode

(vgl. KALBFLEICH/PRENTICE 1980) stehen BRESLOW's Ansatz sowie eine Näherungsformel von EFRON (1977) zur Verfügung. Man kann aber auch ein zeitdiskretes PL-Modell wählen.

#### 5.5.1.4 Die Schätzung der Überlebenswahrscheinlichkeit

Nach den bisherigen Ausführungen dürfte deutlich geworden sein, wie man PL-Schätzungen durchführt und welche Probleme dieser Ansatz hat. Die berechneten Schätzer und ihre Varianz-Kovarianz-Matrix geben Auskunft darüber, welchen Einfluß die einzelnen Variablen haben. Was ist aber, wenn man mit den Ergebnissen Prognosen machen möchte, z.B. um die Schätzungen etwas anschaulicher zu präsentieren. In diesem Fall benötigt man zumindest Angaben über die Überlebenswahrscheinlichkeit. Ihre Berechnung setzt jedoch voraus, daß man die unbekannte Basisrate $\lambda_0(t)$ kennt. Wie man dennoch zu einer Schätzung von S(t) kommt, möchte ich in diesem Abschnitt zeigen.

Die exakte, aber auch die schwierigste Lösung wäre eine Maximierung der Likelihood-Funktion (5.70b), um Schätzer sowohl für die Parameter $\beta$ als auch für S(t) zu erhalten. Zwei alternative Verfahren beruhen auf folgenden Überlegungen:

1. Angenommen, die PL-Schätzer für $\beta$ stimmen mit den wahren ML-Schätzern überein, dann kann man diese Werte in (5.70b) einsetzen und in einem zweiten Schritt ML-Schätzer für S(t) berechnen. Dieser Ansatz wurde von KALBFLEISCH und PRENTICE (1973) vorgeschlagen.
2. Angenommen, die Rate ist zwischen den Ereignissen konstant, kann aber mit jedem Ereignis einen anderen Wert annehmen. In diesem Fall wird der unbekannte Verlauf der Basisrate durch eine Treppenfunktion angenähert. Dieser Ansatz wurde von BRESLOW (1974) vorgeschlagen.

Die erste Methode teilt das ganze Problem in zwei Schritte auf. Die gesuchte Überlebenswahrscheinlichkeit ergibt sich unter der Annahme, daß die Berechnungen des ersten Schrittes (PL-Schätzung) richtig sind. Um dann im zweiten Schritt die Überlebenswahrscheinlichkeit konkret zu berechnen, macht man sich ähnliche Überlegungen zunutze, wie sie für die Ableitung des Kaplan-Meier-Schätzers notwendig sind. Wenn immer

nur ein Ereignis pro Zeitpunkt $t_i$ auftritt, kann man einen solchen *nicht-parametrischen ML-Schätzer der Überlebensfunktion S(t)* folgendermaßen berechnen:

$$(5.73) \qquad \hat{S}_0(t) = \prod_{i|t_i<t} \left[ 1 - \frac{\exp(\mathbf{x}_i\hat{\boldsymbol{\beta}})}{\sum_{l \in R(t_i)} \exp(\mathbf{x}_l\hat{\boldsymbol{\beta}})} \right]^{\exp(-\mathbf{x}_i\hat{\boldsymbol{\beta}})}$$

Sollten allerdings mehrere Ereignisse zum gleichen Zeitpunkt (ties) stattfinden, dann ist die Berechnung etwas komplizierter. Für diesen Fall und die Ableitung von (5.73) verweise ich auf LAWLESS (1982: 359ff.) und KALBFLEISCH/PRENTICE (1980: 84ff.). An dieser Stelle ist es lediglich interessant festzustellen, was passiert, wenn keine Kovariaten berücksichtigt werden. Für $\mathbf{x}_i\boldsymbol{\beta} = 0$ und $d_i = 1$ reduziert sich (5.73) auf:

$$\hat{S}_0(t) = \prod_{i|t_i<t} \left[ 1 - \frac{1}{n_i} \right] = \prod_{i|t_i<t} \frac{n_i - d_i}{n_i}$$

Dieser Ausdruck entspricht genau dem Kaplan-Meier-Schätzer (4.12) der Überlebenswahrscheinlichkeit.

Die zweite Methode ist quasi eine *modifizierte Sterbetafelschätzung*, da der gesamte Prozeß in mehrere Intervalle aufgeteilt wird, die durch die einzelnen Ereignisse definiert werden. Sie verwendet ähnliche Argumente wie ich sie an einigen Stellen in Abschnitt 4 verwendet habe. In diesem Fall kann man zeigen (vgl. z.B. ELANDT-JOHNSON/JOHNSON 1980: 359ff.), daß die unbekannte Basisrate am besten durch folgenden Ausdruck geschätzt wird:

$$(5.74) \qquad \hat{\lambda}_0(t_i) = \frac{d_i}{\Delta t_i \sum_{l \in R(t_i)} \exp(\mathbf{x}_l\hat{\boldsymbol{\beta}})} \qquad \text{mit } \Delta t_i = t_{i+1} - t_i$$

Hieraus ergibt sich dann die Überlebenswahrscheinlichkeit wie folgt:

$$(5.75) \qquad \hat{S}_0(t) = \exp[-\hat{H}_0(t)] \quad \text{mit } \hat{H}_0(t) = \sum_{i|t_i<t} \frac{d_i}{\sum_{l \in R(t_i)} \exp(\mathbf{x}_l\hat{\boldsymbol{\beta}})}$$

Beide Gleichungen reduzieren sich ebenfalls auf (4.16) und (4.17), wenn keine Kovariaten berücksichtigt werden.

*Tabelle 5.4: Schätzung der Überlebenswahrscheinlichkeit für Personen mit und ohne soziale Probleme (KFN−Daten)*

| | Breslow | | | | Kalbfleisch/Prentice | |
|---|---|---|---|---|---|---|
| $t_i$ | $\hat{H}_0(t)$ | $\hat{S}_0(t)$ | $\hat{H}_0(t)^{\exp\hat{\beta}_1}$ | $\hat{S}_0(t)^{\exp\hat{\beta}_1}$ | $\hat{S}_0(t)$ | $\hat{S}_0(t)^{\exp\hat{\beta}_1}$ |
| 0 | 0,000 | 1,000 | 0,000 | 1,000 | 1,000 | 1,000 |
| 2 | 0,027 | 0,973 | 0,153 | 0,858 | 0,971 | 0,847 |
| 3 | 0,067 | 0,935 | 0,379 | 0,685 | 0,928 | 0,656 |
| 4 | 0,088 | 0,916 | 0,495 | 0,609 | 0,908 | 0,581 |
| 5 | 0,099 | 0,906 | 0,558 | 0,572 | 0,898 | 0,546 |
| 6 | 0,121 | 0,886 | 0,685 | 0,504 | 0,878 | 0,480 |
| 7 | 0,144 | 0,866 | 0,814 | 0,443 | 0,857 | 0,419 |
| 10 | 0,172 | 0,842 | 0,974 | 0,378 | 0,833 | 0,357 |
| 12 | 0,221 | 0,802 | 1,246 | 0,288 | 0,791 | 0,267 |
| 14 | 0,294 | 0,745 | 1,660 | 0,190 | 0,728 | 0,166 |
| 16 | 0,393 | 0,675 | 2,221 | 0,109 | 0,650 | 0,088 |
| 17 | 0,426 | 0,653 | 2,405 | 0,090 | 0,629 | 0,073 |
| 21 | 0,462 | 0,630 | 2,609 | 0,074 | 0,606 | 0,059 |

Unter Verwendung der PL−Schätzer $\hat{\beta}$ kann man mit der Überlebenswahrscheinlichkeit (5.73) oder (5.75) Prognosen für jede Konstellation der Kovariaten machen. Dieses Vorgehen ist in Tabelle 5.4 illustriert, in der mit den Daten aus Tabelle 4.4 die Überlebensfunktion der Personen mit und ohne soziale Probleme berechnet wurde (PRG0502). Wegen der ties bietet sich für eine Berechnung mit dem Taschenrechner der Ansatz von BRESLOW an. Bei einem Regressionsmodell mit dichotomer Kovariaten (vgl. Gleichung 5.64) reduziert sich der Ausdruck (5.75) für die geschätzte kumulierte Rate auf

$$\hat{H}_0(t) = \sum_{i|t_i<t} \frac{d_i}{n_{i1}\exp\hat{\beta}_1 + n_{i2}},$$

so daß mit Hilfe des PL−Schätzers $\hat{\beta}_1 = 1{,}7316$ (s.oben) sowie der Ereignisse und Risikomengen aus Tabelle 4.4 $H_0(t)$ bzw. $S_0(t) = \exp[-H_0(t)]$ leicht berechnet werden können. Die Ergebnisse finden sich in Tabelle 5.4 unter der Überschrift "Breslow". $\hat{S}_0(t)$ entspricht der geschätzten Überlebenswahrscheinlichkeit der Personen ohne soziale Probleme. Die entsprechende Überlebenswahrscheinlichkeit der Personen mit sozialen Problemen

entspricht gemäß (5.4b) $\hat{S}_0(t)^{\exp\beta_1}$. Der Vollständigkeit halber enthält Tabelle 5.4 auch die entsprechenden Schätzungen nach dem Vorschlag von KALBFLEISCH und PRENTICE, die auf Grund der ties iterativ mit einem Computer-Programm bestimmt werden müssen. Die numerischen Unterschiede zwischen beiden Schätzern sind gering.

### *5.5.2 Eine Illustration des Regressionsmodells von Cox*

Um einen Eindruck davon zu vermitteln, wie man konkret eine PL-Analyse durchführt und interpretiert, habe ich das Rückfallrisiko von Strafentlassenen nach verschiedenen Problemfeldern untersucht (Abschnitt 5.5.2.1). Bei der Interpretation der Ergebnisse ergeben sich keine wesentlichen Unterschiede gegenüber dem bisherigen Vorgehen, weder inhaltlich noch methodisch. Ich werde mich daher hauptsächlich mit der Frage beschäftigen, wie man die zentrale Annahme dieses Modells, die Existenz proportionaler Risiken testen kann (Abschnitt 5.5.2.2).

#### 5.5.2.1 Schätzung von Modellen mit partiell spezifizierter Rate

Die Ergebnisse der PL-Schätzung für die KFN-Daten zeigt Tabelle 5.5 (PRG0502). Parameter, Standardfehler und Antilogarithmen der einzelnen Effekte können wie bisher interpretiert werden (vgl. Abschnitt 5.3). Man beachte, daß aus den genannten Gründen eine Regressionskonstante nicht geschätzt werden kann.

Modell 1 zeigt noch einmal den bereits bekannten Effekt sozialer Probleme. In Modell 2 werden zusätzlich die weiteren Belastungen der Strafentlassenen (Termine, Schulden, Arbeit, Wohnung) berücksichtigt. Bis auf den Effekt der Schulden haben alle Parameter positive Vorzeichen, erhöhen also das Rückfallrisiko. Bei Berücksichtigung der Standardfehler zeigt sich jedoch, daß keiner dieser Effekte signifikant ist. Die Berücksichtigung zusätzlicher Problemfelder verbessert also Modell 1 nur unwesentlich[1], was sicherlich auch ein Resultat der kleinen Fallzahl ist. Zur anschaulichen Interpretation verwendet man am besten die Antilogarithmen. Die Signifikanz der einzelnen Parameter läßt sich, wie vorher auch, mit Hilfe der

1) Der partielle Likelihood-Verhältnis-Test ist mit $X^2 = 4{,}68$ bei 4 Freiheitsgraden nicht signifikant.

geschätzten Varianz-Kovarianz-Matrix testen. Gruppen von Variablen sowie der Modellfit insgesamt können mit Hilfe des Likelihood-Verhältnis-Testes untersucht werden.

Man sieht also, daß sich am bisherigen Vorgehen nichts ändert, weil die partielle Likelihood-Funktion behandelt wird wie eine Likelihood-Funktion im üblichen Sinne. Ein Nullmodell in dem o.g. Sinne (nur eine Re-

*Tabelle 5.5: Modelle mit partiell spezifizierter Rate (MZ71-Daten)*

| Variable | | Modell 1 | Modell 2 | Modell 3 | Modell 4 |
|---|---|---|---|---|---|
| Soziale | a) | 1,7316* | 1,3394* | | 1,1740* |
| Probleme | b) | 0,3753 | 0,4610 | | 0,5355 |
| | c) | 5,6494 | 3,8166 | | 3,2349 |
| Interaktion | | | | | -1,1229 |
| | | | | | 0,6543 |
| | | | | | 0,3253 |
| Termine | | | 0,1680 | 0,1331 | |
| | | | 0,5365 | 0,5463 | |
| | | | 1,1829 | 1,1424 | |
| Schulden | | | -0,2992 | -0,2993 | |
| | | | 0,3942 | 0,3976 | |
| | | | 0,7414 | 0,7413 | |
| Arbeit | | | 0,4890 | 0,4614 | |
| | | | 0,5091 | 0,5171 | |
| | | | 1,6307 | 1,5863 | |
| Wohnung | | | 0,9537 | 0,5712 | |
| | | | 0,7073 | 0,7141 | |
| | | | 2,5954 | 1,7703 | |
| N | | 64 | 64 | 64 | 64 |
| D | | 31 | 31 | 31 | 31 |
| in % | | 48,4% | 48,4% | 48,4% | 48,4% |
| ln L (Modell 0) | | -117,80 | -117,80 | -90,67 | -117,80 |
| ln L | | -108,16 | -105,82 | -89,24 | -106,34 |
| Pseudo-$R^2$ | | 8,2% | 10,2% | 1,6% | 9,7% |
| L-Ratio | | 19,27 | 23,95 | 2,86 | 22,90 |
| df | | 1 | 5 | 4 | 2 |

a) Parameter (* sign. 5%-Niveau), b) Standardfehler, c) Antilogarithmus

gressionskonstante) gibt es jedoch nicht, da das Cox'sche Regressionsmodell keine Konstante enthält. Man kann sich jedoch leicht überlegen, daß die partielle Log-Likelihood (5.65) $\sum_t d_t \ln(n_{t.})$ entspricht, wenn keine Kovariaten berücksichtigt werden. Auf diese Weise kann man die Likelihood eines Nullmodells bestimmen und die in Abschnitt 5.3 diskutierten globalen Likelihood-Verhältnis-Tests und Pseudo-$R^2$-Werte berechnen (vgl. Tabelle 5.5).

Da die partielle Likelihood-Funktion nur untersucht, welche Personen bei einer gegebenen Risikomenge ausscheiden, wird üblicherweise behauptet, daß sich die Verhältnisse zum Zeitpunkt des Ereignisses besonders gut berücksichtigen lassen. Man könnte also *zeitabhängige Variablen* bequem mit ihren Werten zum Zeitpunkt des Wechsels berücksichtigen. Die entsprechende Kodierungsvorschrift würde etwa lauten: Verwende bei zeitabhängigen Merkmalen immer die Werte, die am Ende eines Zustands gelten. So einfach ist das Problem jedoch nicht zu lösen, denn damit hat man lediglich den Zähler der Likelihood-Funktion (5.71) aktualisiert. Im Nenner stehen aber noch Individuen, die zur Risikomenge gehören. Nach der obigen Kodierungsvorschrift hätten sie bei den zeitabhängigen Merkmalen Werte, die erst für spätere Zeitpunkte gelten, wenn nämlich diese Individuen ein Ereignis haben. Richtig wäre also eine Prozedur, die für jeden Zeitpunkt eines Ereignisses die Werte der zeitabhängigen exogenen Merkmale für alle risikobelasteten Individuen aktualisiert. Ein solches zeitabhängiges Modell werde ich im folgenden Abschnitt verwenden.

#### 5.5.2.2 Geschichtete PL-Modelle und Test der Annahme proportionaler Risiken

Alle bisher besprochenen Modelle gehen von einem proportionalen Risiko und log-linearen Abhängigkeiten aus. Die Raten zweier beliebiger Personen unterscheiden sich also nur durch einen Proportionalitätsfaktor und der Effekt der Kovariaten ist multiplikativ. Das COX'sche Regressionsmodell macht ganz besonders von dieser Annahme Gebrauch, denn eine Eliminierung der Basisrate ist bei einem additiven Modell nicht möglich (vgl. Gleichung 5.63). Wie kann man diese Annahme überprüfen und wie läßt sich das Modell verallgemeinern, wenn die Annahme nicht gegeben ist? Beginnen wir mit der Verallgemeinerung.

Man erinnere sich daran, daß die Basisrate $\lambda_0(t)$ für alle Personen gleich ist, so daß man lediglich eine Basis – Überlebenswahrscheinlichkeit kennen muß, um das Überleben einer beliebigen Person zu prognostizieren. Diese individuellen Überlebensfunktionen werden sich nicht überkreuzen (vgl. Gleichung 5.5). Dieses Modell ist natürlich sehr starr. Es kann jedoch flexibler gemacht werden, wenn man verschiedene Gruppen $q=1,\ldots,Q$ definiert, zwischen denen diese Basis – Überlebenswahrscheinlichkeit variieren darf. Das impliziert gleichzeitig eine *gruppenspezifische Basisrate* $\lambda_0^q(t)$:

$$(5.76) \qquad S_q(t) = [S_0^q(t)]^{\exp(\mathbf{x}_i\beta)} \quad \text{mit } S_0^q(t) = \exp[-\int_0^t \lambda_0^q(u)du]$$

Für jede dieser Gruppen kann man nun ein eigenes Regressionsmodell schätzen und im allgemeinen können sich die geschätzten Überlebensfunktionen überschneiden, wenn man sie alle in ein gemeinsames Koordinatenkreuz einzeichnet.

Man beachte, daß bei einer getrennten Auswertung für jede Gruppe die Effekte $\beta$ ebenfalls von Schätzung zu Schätzung variieren können (Disaggregierung). Wenn man aber nur die Basisrate variieren lassen möchte, dann muß man eine Auswertung mit allen Fällen vornehmen, bei der die Basisrate frei zwischen den Gruppen variieren darf, die Parameter $\beta$ aber einheitlich geschätzt werden (*Schichtung*). Eine solche Schätzung ergibt sich, wenn man das Produkt aller gruppenspezifischen Likelihood – Funktionen betrachtet. Bei der (partiellen) Likelihood – Funktion (5.71) muß man dazu noch einen zusätzlichen Term für alle Gruppen betrachten:

$$(5.77) \qquad L = \prod_q^Q \prod_j^J \prod_k^K \prod_\iota^I \left( \frac{\mathbf{x}_{\iota jk}\boldsymbol{\beta}_{jk}}{\sum_{l \in R_j^q(t_\iota)} \exp(\mathbf{x}_{ljk}\boldsymbol{\beta}_{jk})} \right)^{\delta_{\iota jk}^q}$$

Auch diese Likelihood – Funktion läßt sich einfach lösen. Die Risikomenge $R_j^q(t_\iota)$ bezieht sich jetzt nur auf die Untersuchungseinheiten im Ausgangszustand j, die zu einer Gruppe q gehören. Vergleiche müssen jetzt nur innerhalb der einzelnen Gruppen durchgeführt werden und die Berechnung

vereinfacht sich weiter[1]. Gruppenspezifische Überlebensfunktionen lassen sich dann mit ähnlichen Methoden schätzen wie ich sie in Abschnitt 5.5.1.4 beschrieben habe.

Mit dieser Erweiterung des Modells kann man die *Annahme proportionaler Risiken* überprüfen, während gleichzeitig mehrere Kovariaten kontrolliert werden. Definiert man die Gruppen an Hand eines Merkmals, das einen multiplikativen Effekt auf die Basisrate hat, dann bringt eine geschichtete Auswertung keinen zusätzlichen Informationsgewinn, denn die gruppenspezifischen Überlebensfunktionen werden sich nur durch einen Proportionalitätsfaktor unterscheiden. Anders hingegen bei Merkmalen, von denen kein multiplikativer Einfluß ausgeht. Das erkennt man leicht durch folgende Überlegung: Angenommen, das in Frage stehende Merkmal $x_p$ habe lediglich zwei Ausprägungen 0 und 1 und der Vektor $\mathbf{x}_t$ enthalte die anderen Kovariaten. Hat $x_p$ einen multiplikativen Einfluß, dann kann man mit (5.4) die gruppenspezifischen Überlebensfunktionen leicht angeben:

$$S_1(t) = S_0(t)^{\exp(\mathbf{x}_t\beta)} \qquad \text{(Gruppe 1: } x_{tp} = 0\text{)}$$

$$S_2(t) = S_0(t)^{\exp\beta_p \exp(\mathbf{x}_t\beta)} \qquad \text{(Gruppe 2: } x_{tp} = 1\text{)}$$

Betrachtet man jetzt eine doppelt logarithmierte Transformation (5.5) der Überlebenswahrscheinlichkeiten, dann erkennt man, daß sich beide nur durch die Konstante $\beta_p$, den Effekt der Variablen $x_p$, unterscheiden:

$$\ln [-\ln S_1(t)] = \mathbf{x}_t\beta + \ln [-\ln S_0(t)]$$

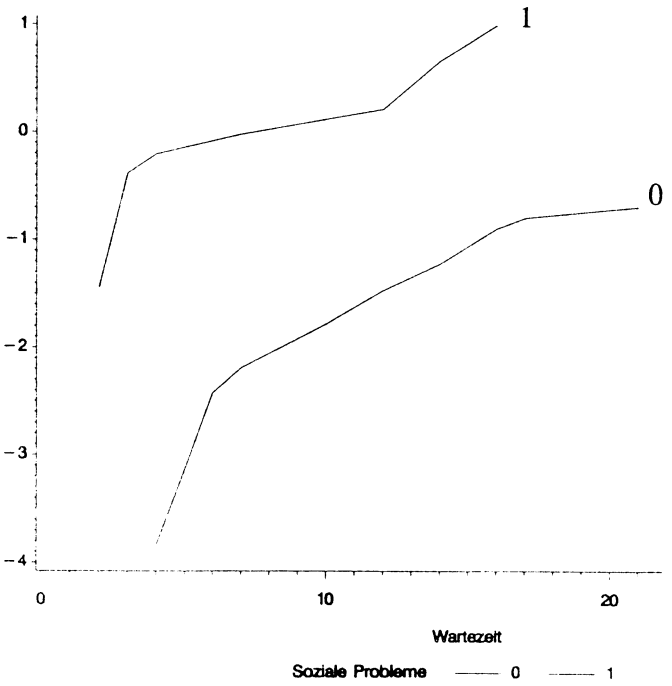
$$\ln [-\ln S_2(t)] = \beta_p + \mathbf{x}_t\beta + \ln [-\ln S_0(t)]$$

Eine Zeichnung dieser doppelt transformierten Überlebenswahrscheinlichkeiten muß also zwei parallele Kurven im Abstand $\beta_p$ ergeben. Hat das Merkmal $x_p$ dagegen keinen multiplikativen Einfluß, dann wird dieser Abstand variieren, so daß sich die beiden Kurven im Extremfall überschneiden können. Durch Schichtung mit der in Frage stehenden Variablen bei

1) Die einfache Rechnung ermöglicht es, schnell einmal den Einfluß einer bestimmten Variablen (z.B. eines bestimmten Treatments) zu testen, während der Einfluß aller anderen Merkmale kontrolliert wird. Man verwendet dazu die interessierende Variable als Gruppierungskriterium.

gleichzeitiger Konstanthaltung aller anderen Variablen kann man also proportionale Risiken prüfen.

*Abbildung 5.2: Überlebensfunktionen des geschichteten Regressionsmodells 3 (KFN – Daten)*




Zur Illustration habe ich mit Hilfe eines solchen *geschichteten Regressionsmodells* geprüft, ob sich die Rückfallrisiken der Personen mit und ohne soziale Probleme proportional verhalten. Gleichzeitig wurden alle weiteren Belastungsfaktoren kontrolliert (Modell 3 in Tabelle 5.5, PRG0502). Auf der Basis dieses Modell wurde die Überlebenswahrscheinlichkeit beider Gruppen geschätzt. Abbildung 5.2 zeigt den Verlauf der doppelt logarithmierten Überlebensfunktion. Da sie näherungsweise in konstantem Abstand zueinander verlaufen, scheint die Annahme proportionaler Risiken gerechtfertigt.

Mit Hilfe einer zeitabhängigen Variablen kann man schließlich inferenzstatistisch prüfen, ob der Abstand über den gesamten Zeitraum konstant ist.

Wenn die Annahme proportionaler Risiken zutreffen soll, dann unterscheiden sich die Risiken zweier Personen nur durch einen Proportionalitätsfaktor, und zwar unabhängig vom Zeitpunkt. Anders ausgedrückt, es darf keinen signifikanten Interaktionseffekt zwischen der in Frage stehenden Variablen und der Zeit geben. Möchte man also prüfen, ob sich die Rückfallrisiken von Personen mit und ohne soziale Probleme proportional zueinander verhalten, definiert man eine Interaktionsvariable Z wie folgt: $z_i = x_{ip} lnt_i^*$ mit $x_{ip} = 0$ (ohne Probleme), $x_{ip} = 1$ (mit Problemen) und $lnt_i^* = \ln(t_i - c)$.[1] Die Raten lauten dementsprechend:

$$r_a(t) = \lambda_0(t) \exp(\mathbf{x}_i\beta) \qquad \text{(ohne Problem)}$$

$$r_b(t) = \lambda_0(t) \exp(\mathbf{x}_i\beta + \beta_p + \gamma z_i) \qquad \text{(mit Problemen)}$$

Sollen sie sich nur durch einen Proportionalitätsfaktor $\exp(\beta_p)$ unterscheiden, dann darf der Effekt $\gamma$ der Interaktionsvariablen nicht signifikant von Null verschieden sein.

Modell 4 enthält eine solche zeitabhängige Interaktionsvariable (vgl. Tabelle 5.5). Ihre Werte müssen für jeden Ereigniszeitpunkt $\iota$ und für jede Beobachtung aktualisiert werden, wenn der Schätzlogarithmus die Risikomenge $R(t)$ im Zeitverlauf abarbeitet. Dies ist in den meisten Programmen für das Cox-Modell vorgesehen (vgl. PRG0502). Der geschätzte Effekt $\hat{\gamma} = -1{,}1229$ ist nicht signifikant und bestätigt damit die Annahme proportionaler Risiken.

## 5.6 *Auswahl und Evaluation eines geeigneten Modells*

In den vorhergehenden Abschnitten habe ich mich mit der Frage beschäftigt, wie man multivariate Regressionsmodelle mit unterschiedlichen Formen der Zeitabhängigkeit spezifiziert, schätzt und interpretiert. Wie Abbildung 3.4 zeigt, ist die Suche nach dem passenden Modell für die untersuchten Daten ein iterativer Prozeß. Am Ende dieses langen Textes ist es daher

1) Wenn die Variable z gleich $x_p \cdot lnt$ wäre, würden die beiden Variablen z und $x_p$ hoch miteinander korrelieren. In diesem Fall ist die Maximierung der Likelihood-Funktion schwierig. Man verwendet daher $lnt^* = \ln(t - c)$, wobei c ungefähr dem Durchschnitt der logarithmierten Wartezeiten entspricht (bei den KFN-Daten also $\ln 11{,}27 = 2{,}422$).

sinnvoll, die vielen Einzelergebnisse noch einmal in bezug auf diese strategische Frage zuzuspitzen.

Bevor man sich für ein konkretes Regressionsmodell entscheidet, sollte man zunächst die Modellannahmen (proportionale Risiken) testen. Dann geht es um die Frage, ob es sich um einen zeitabhängigen Prozeß handelt und wenn ja um welchen. Das setzt voraus, daß man wesentliche Einflußfaktoren berücksichtigt (multivariates Modell), denn unkontrollierte Heterogenität erscheint als negative Zeitabhängigkeit (vgl. Abschnitt 2.5.1). Diese *schrittweise Modellauswahl* kann bei einer konkreten Datenanalyse folgendermaßen durchgeführt werden:

1. Auswahl geeigneter Kovariaten mit Hilfe nicht-parametrischer Verfahren.
2. Ergeben sich dabei proportionale Risiken?
3. Test multivariater proportionaler Risiken durch das partiell parametrische Modell von COX.
4. Entscheidung für ein zeitabhängiges Modell mit Hilfe der Modellprognosen aus Schritt 3.

Ich will diese vier Analyseschritte im folgenden kurz erläutern.

Wie man den Einfluß (weniger) diskreter Kovariaten mit Hilfe nicht-parametrischer Verfahren untersucht, habe ich in Abschnitt 4 ausführlich dargestellt. Auf dieser Stufe der Datenanalyse können metrische Variablen nur nach vorheriger Gruppierung berücksichtigt werden, was in dieser explorativen Phase kein wesentlicher Informationsverlust ist. Außerdem kann man hier schon die Annahme proportionaler Risiken testen, denn für die doppelt logarithmierte Überlebenswahrscheinlichkeit verschiedener Gruppen gilt, daß sie sich nur durch einen konstanten Faktor unterscheiden dürfen (vgl. Gleichung 5.5). Die Überlebensfunktionen $\ln(-\ln(S(t))$ verschiedener Gruppen müssen also im konstanten Abstand parallel zueinander verlaufen und dürfen sich nicht überschneiden, wenn das Modell proportionaler Risiken richtig sein soll.

Leider ist die Zahl der simultan analysierbaren Merkmale mit diesem Ansatz begrenzt. An dieser Stelle bietet sich das partiell parametrische Regressionsmodell von COX an. Mit ihm kann man mehrere Kovariaten unterschiedlichen Meßniveaus kontrollieren, während die Annahme proportionaler Risiken für eine unabhängige Variable getestet wird. Dazu wird diese Variable als Schichtungskriterium verwendet und das Regressionsmodell mit den verbleibenden Kovariaten wird simultan in den verschiede-

nen Schichten geschätzt (vgl. Abschnitt 5.5.2.2). Auch hier müssen die Überlebensfunktionen $\ln(-\ln(S(t))$ der verschiedenen Schichten parallel verlaufen, wenn die Modellannahme richtig ist. Im Gegensatz zu den einfacheren explorativen Methoden aus Abschnitt 4 können hier jedoch multivariate Zusammenhänge und metrische Merkmale berücksichtigt werden.

Aus dem Verlauf der Überlebensfunktionen kann man dann in einem weiteren Schritt bestimmte zeitabhängige Regressionsmodelle ableiten. Wenn beispielsweise eine Zeichnung der doppelt logarithmierten Überlebensfunktion $\ln(-\ln S(t))$ (y-Achse) mit $\ln t$ (x-Achse) eine Gerade ergibt, dann handelt es sich um eine Weibull-Verteilung. Ähnliche graphische Tests für andere Verteilungsmodelle habe ich in Kapitel 4 besprochen (vgl. Gleichung 4.29). Für S(t) verwendet man am besten die Ergebnisse des ungeschichteten Regressionsmodells. Mit den PL-Schätzern kann man S(t) für jede Konstellation der Kovariaten berechnen, z.B. für den Durchschnitt jedes Merkmals (vgl. Abschnitt 5.5.1.4).

Der Vorteil des COX'schen Regressionsmodells besteht darin, daß es möglichst viele Kovariaten kontrolliert, ohne den zeitlichen Verlauf der Rate festzulegen. Von daher werden scheinbare Zeitabhängigkeiten durch unkontrollierte Heterogenität weitgehend ausgeschlossen und man kann verschiedene zeitabhängige Modelle durch Plots der Überlebensfunktion testen, deren Verlauf im Modell zunächst einmal nicht festgelegt wird.

Angenommen man hat sich für ein konkretes Regressionsmodell entschieden und dessen Parameter mit Hilfe empirischer Daten geschätzt. Jetzt stellt sich die Frage, welche Schlußfolgerungen (Prognosen) aus den Ergebnissen gezogen werden können und wie gut das Modell auf die Daten paßt. In Abschnitt 5.3.1.3 und 5.5.1.4 habe ich gezeigt, wie man mit den geschätzten Parametern für beliebige Variablenkonstellationen die Überlebenswahrscheinlichkeit einzelner Gruppen berechnen kann. Diese Modellprognosen kann man mit nicht-parametrischen Schätzungen der Überlebenswahrscheinlichkeit vergleichen und erhält auf diese Weise erste Aufschlüsse über die Gruppen, deren Verhalten durch das Modell nur unzureichend beschrieben werden kann. Ein Beispiel für diese Strategie findet sich bei TUMA et al. (1979: 843).

Dieses Vorgehen ist aber nicht besonders hilfreich, da es immer eine Disaggregierung des Datenmaterials voraussetzt. Möchte man jede Konstellation der unabhängigen Merkmale betrachten, dann wächst die Zahl der zu analysierenden Gruppen sehr schnell ins Unüberschaubare. Bei metrischen Kovariaten kann eine Gruppe im Extremfall sogar aus nur einer Beobach-

tung bestehen. Daher wäre es wünschenwert, wenn man, wie bei klassischen Regressionsmodellen, für jeden Fall ein Residuum berechnen könnte (vgl. Abschnitt 6.5).

# 6. Weiterführende Fragestellungen

Es liegt auf der Hand, daß in einem einführenden Text nicht alle Details und Weiterentwicklungen der Verlaufsdatenanalyse angesprochen werden können. Abschließend möchte ich daher mit einigen praktischen Hinweisen und Literaturangaben auf einige weiterführende Fragestellungen aufmerksam machen.

## 6.1 Regressionsmodelle mit zeitabhängigen Kovariaten

In Abschnitt 3.3.3 habe ich zeitabhängige Kovariaten unter datentechnischen Gesichtspunkten diskutiert. An dieser Stelle möchte ich einige theoretische Ergänzungen und praktische datenanalytische Hinweise geben. Dabei sei jedoch schon jetzt darauf hingewiesen, daß die Problematik zeitabhängiger Merkmale und die daraus resultierenden Fehlspezifikationen bei weitem noch nicht erforscht sind. Weitere statistische Hintergrundinformationen erhält man in Kapitel 5 des Buches von KALBFLEISCH und PRENTICE (1980).

Zunächst ist es hilfreich, die verschiedenen Arten von zeitabhängigen Merkmalen begrifflich voneinander zu trennen. KALBFLEISCH und PRENTICE unterscheiden in diesem Zusammenhang zwischen *a) externen und b) internen Variablen* (1980: 122ff.). Bei ersteren sind alle Werte bekannt, ohne daß man den Prozeß selbst kennen muß, während bei letzteren die Beobachtung einer konkreten Ausprägung von den Ergebnissen des Prozesses selbst abhängt. Betrachtet man zur Illustration die MZ71 – Daten, dann wären der Status des Vaters oder die Einberufung zur Bundeswehr Beispiele für externe Variablen, während die Entscheidung für eine Weiterbildung von dem konkreten Berufsverlauf abhängt.

Bei den externen Merkmalen unterscheiden KALBFLEISCH und PRENTICE weiterhin zwischen *fixierten, definierten und stochastischen Variablen*:

aa) fixierte Variablen – zeitkonstante oder als zeitkonstant behandelte Merkmale, z.B. Geschlecht oder Status des Vaters.

ab) definierte Variablen – zeitabhängige Merkmale mit festgelegtem Verlauf, z.B. Berufserfahrung.

ac) stochastische Variablen – Ergebnisse eines anderen (unabhängigen) stochastischen Prozesses, z.B. Wehrdienst[1].

Interne Variablen in dem oben definierten Sinn sind die Merkmale "Weiterbildung", "Beschäftigtenwachstum Zielbranche", "Tätigkeitswechsel 1967" und "Nummer der Tätigkeit". Auch hier ist es für die folgende Darstellung hilfreich, weiter zu differenzieren:

ba) einfacher Fall – Um einen bestimmten Wert der Variablen zu beobachten, muß die Person im Ausgangszustand verbleiben (vgl. Weiterbildung: Um eine Unterbrechung zu registrieren, muß die Tätigkeit weiter ausgeübt werden.).

bb) komplizierter Fall – Um einen bestimmten Wert zu beobachten, muß ein Ereignis stattfinden (vgl. die drei anderen Merkmale: Um hier einen Wert beobachten zu können, muß ein Tätigkeitswechsel vorliegen).

KALBFLEISCH/PRENTICE (1980: 122ff.) haben gezeigt, daß für die zeitabhängigen Variablen ab – ba eine allgemeine Likelihood – Funktion existiert, die es erlaubt, den Einfluß dieser Merkmale exakt zu schätzen. Diese theoretisch – statistische Ableitung überschreitet den Rahmen dieser Abhandlung. Ich will mich daher mit den datenanalytischen Konsequenzen dieser Überlegungen beschäftigen.

In allen Auswertungen mit den MZ71 – Daten habe ich die Kovariaten mehr oder weniger als zeitkonstant betrachtet (fixierte Variablen). Die einzige Ausnahme war die Variable "Berufserfahrung" (definierte Variablen). In diesem Fall ergab sich ein zeitabhängiges Regressionsmodell. Oder anders ausgedrückt, die Zeit t einer Gompertz – oder Wiebull – Rate erhielt eine substantielle Interpretation und entsprach der Zeitdauer seit Berufseintritt (= Berufserfahrung). Definierte Variablen sind daher abgeleitete Größen der Prozeßzeit und ergeben die bekannten zeitabhängigen Regressionsmodelle.

Eine angemessene Auswertung zeitabhängiger Merkmale muß Art und Zeitpunkt der Veränderungen berücksichtigen. Alle genannten Likelihood – Funktionen für Regressionsmodelle mit Raten bestehen mindestens aus zwei Teilen: Erstens aus der Überlebenswahrscheinlichkeit, daß bis zu

1) Dabei wird angenommen, daß die Einberufung durch die Kreiswehrersatzämter vorgenommen wird und die Person den Wehrdienst weitgehend unabhängig von den eigenen Karriereentscheidungen ableisten muß.

einem bestimmten Zeitpunkt keine Ereignisse auftreten, und zweitens aus der Rate, daß dann zu diesem Zeitpunkt Veränderungen stattfinden. Genauso wie die Überlebenswahrscheinlichkeit die Veränderung der Rate mit der Zeit t berücksichtigt (vgl. Gleichung 2.13), müßte sie auch die Veränderung mit den zeitabhängigen Kovariaten erfassen. Sind diese Merkmale mehr oder weniger mit der Zeit t identisch (definierte Variablen), dann sind die zeitabhängigen Modelle aus Abschnitt 5.4 schon die Lösung. In allen anderen Fällen müßte jedoch der zeitliche Verlauf der Kovariaten in der Überlebenswahrscheinlichkeit berücksichtigt werden.

Bei den *partiell parametrischen Methoden* ist die Situation besonders einfach, da hier lediglich die Verhältnisse zum Zeitpunkt eines Ereignisses berücksichtigt werden. Von daher muß man lediglich für jedes Ereignis die Risikomenge aktualisieren. Das ist mit allen Programmen für das Cox'sche Regressionsmodell (vgl. Anhang D und PRG0502) leicht möglich: Angenommen man kennt den Zeitpunkt der Einberufung und hat ihn in einer Hilfsvariablen EINBERUF abgespeichert. Durch Vergleich mit der Prozeßzeit t erhält die eigentlich interessierende Variable BUNDWEHR erst zum eigentlichen Einberufungszeitpunkt den Wert 1 (vorher hat sie den Wert 0).

Da bei den (vollständig) parametrischen Modellen eine solche einfache Programmlösung nicht besteht, muß man sich hier anderweitig weiterhelfen. Eine Möglichkeit ist das sogenannte *Episodensplitting* (BLOSSFELD et al. 1986: 193ff.). Dabei wird eine Beobachtung in mehrere Teilbeobachtungen zerlegt, so daß nicht das Problem entsteht, daß man nur einen Wert der zeitabhängigen Variablen pro Datensatz berücksichtigen kann. Wenn man immer dann einen Teil der ursprünglichen Beobachtung absplittet, wenn sich das zeitabhängige Merkmal verändert, dann kann man im Prinzip alle im Zeitablauf auftretenden Werte berücksichtigen[1]. Dabei muß man aber eine immense Erhöhung der Fallzahl in Kauf nehmen.

Eine andere Möglichkeit ist das in Abschnitt 5.4.2 besprochene *periodenspezifische Modell*, das sich mit dem Programm *RATE* leicht realisieren läßt (und zwar für zeitkonstante wie für zeitabhängige Raten). Da man für jede Periode einen neuen Set von Kovariaten angeben kann, kann man

1) Vgl. unsere Überlegungen zur Kodierung zeitabhängiger Variablen bei zeitdiskreten Daten (Abschnitt 3.3.3). Auf diese Weise ist auch die gegenseitige Beeinflussung von mehreren Variablen modellierbar (multidimensionaler Zustandsraum). Vgl. z.B. SÖRENSEN/SÖRENSEN (1983) über den Zusammenhang von Schulbesuch und Wohnen im Elternhaus.

natürlich auch die aktualisierten Werte der zeitveränderlichen Variablen einsetzen. Im Gegensatz zu der vorher besprochenen Splittingstrategie erhöht sich hier die Anzahl der zu schätzenden Parameter und nicht die Fallzahl. Dort wird unter Berücksichtigung aller im Zeitablauf auftretenden Werte ein Durchschnittseffekt berechnet, während hier genauso viele Effekte wie Perioden geschätzt werden. Außerdem sind die Variablen, die nur für jede Periode aktualisiert werden, hochgradig multikollinear.

Eine besondere Schwierigkeit sind schließlich zeitabhängige Merkmale des Typs bb, die direkt mit dem Auftreten eines Ereignisses verbunden sind. Da sie die Art des Ereignisses beschreiben, definieren sie eigentlich einen neuen Zustandsraum. Bei allen Variablen des Typs bb ist daher zu prüfen, ob es nicht eine *andere Definition des Zustandsraumes* gibt, die diese Merkmale berücksichtigt. Ansonsten verbleibt nur die Möglichkeit, sogenannte *Indikatormerkmale* zu verwenden. Statt der Variablen "Tätigkeitswechsel 1967" also die Prozeßzeit t.

## *6.2 Regressionsmodelle mit unbeobachteter Heterogenität*

Ein ganz einfaches Modell mit unbekannter Heterogenität ergibt sich, wenn man das Modell exponentiell verteilter Wartezeiten um einen Fehlerterm $\epsilon$ erweitert (vgl. Gleichung 5.2d):

$$(6.1) \qquad r(t|\mathbf{x}_i,\epsilon) = \exp(\mathbf{x}_i\boldsymbol{\beta} + \epsilon^*) \quad \text{mit } \exp\epsilon^* = \epsilon$$

Diese Fehler $\epsilon = \exp\epsilon^*$ seien gamma-verteilt mit Mittelwert 1 und Varianz $\sigma^2$.[1] Für die Analyse multipler und wiederholbarer Ereignisse wird zunächst davon ausgegangen, daß die Fehler $\epsilon$ für verschiedene Risiken und Ereignisse nicht miteinander korrelieren.

Wenn keine Kovariaten berücksichtigt werden und praktisch nur unbekannte Heterogenität auftritt, reduziert sich (6.1) auf:

$$(6.2) \qquad r(t|\epsilon_i) = \epsilon_i$$

1) Die Gamma-Verteilung hat den Vorteil großer Flexibilität, da man die unterschiedlichsten Verteilungsformen (symmetrisch, linksschief, rechtsschief) modellieren kann. Da der Fehlerterm multiplikativ wirkt, muß er einen Mittelwert von 1 haben.

D.h. jede Person hat eine konstante aber verschiedene Rate. Die Verteilung dieser Raten wird durch die Gamma–Verteilung beschrieben. Wartezeiten sind jetzt nur noch auf der individuellen Ebene exponentiell verteilt:

$$f(t_i|\epsilon_i) = \epsilon_i \exp(-\epsilon_i t_i) \tag{6.3}$$

Für die gesamte Untersuchungsgruppe muß man quasi alle individuellen Verteilungen zusammenrechnen. Das geschieht durch Integration:

$$f(t) = \int_0^\infty f(t|\epsilon)\, g(\epsilon)\, d(\epsilon) = \frac{\gamma\, \lambda^\gamma}{(\lambda + 1)^{\gamma+1}} \tag{6.4}$$

Dabei entspricht $g(\epsilon)$ der Dichte der Gamma–Verteilung (s. Anhang B). Auf der Ebene der gesamten Untersuchungsgruppe ergibt sich also eine andere Verteilung der Wartezeiten als auf der individuellen Ebene. Dieses Modell ist auch als *Yule–Greenwood–Prozeß* bekannt. Da hier verschiedene individuelle Verteilungen zusammengefaßt werden, verwendet man auch den Begriff *Mischverteilung*. Weiterhin kann man zeigen, daß die Gesamtrate eines Yule–Greenwood–Prozesses folgendem Ausdruck entspricht:

$$r(t) = \frac{\gamma}{\lambda + t} \tag{6.5}$$

Es ergibt sich eine im Zeitablauf fallende Rate (Hyperbel). Dieses Ergebnis deckt sich mit den Überlegungen in Abschnitt 2.5.1, wonach unkontrollierte Heterogenität scheinbare negative Zeitabhängigkeit ergibt. Dies impliziert, daß ein Yule–Greenwood–Prozeß zunächst nicht von einem zeitabhängigen Modell zu unterscheiden ist.

Natürlich kann man auch bekannte Heterogenitätsfaktoren berücksichtigen (wie in Gleichung 6.1 vorgesehen). Es läßt sich eine Likelihood–Funktion angeben und mit Hilfe einer ML–Schätzung kann man die unbekannten Parameter $\beta$ und die Varianz der Fehler $\epsilon$ bestimmen. Dieser Vorschlag stammt von TUMA (1982a), die dieses Modell auf eine Analyse von Tätigkeitswechseln angewendet hat. DIEKMANN (1984) diskutiert einige praktische Implikationen.

*Tabelle 6.1: Modelle mit unbeobachteter Heterogenität (KFN – Daten)*

| Variable | | Modell 0 | Modell 1 | Modell 2 |
|---|---|---|---|---|
| Konstante | a) | −3,1253* | −4,0331* | −3,9934* |
| | b) | 0,3118 | 0,2796 | 0,3333 |
| | c) | 0,0439 | 0,0177 | 0,0184 |
| Termine | | | | 0,1836 |
| | | | | 0,5317 |
| | | | | 1,2015 |
| Schulden | | | | −0,3106 |
| | | | | 0,3851 |
| | | | | 0,7330 |
| Arbeit | | | | 0,6212 |
| | | | | 0,5047 |
| | | | | 1,8612 |
| Soziale | | | 2,0016* | 1,4177* |
| Probleme | | | 0,4226 | 0,4490 |
| | | | 7,4009 | 4,1276 |
| Wohnung | | | | 0,7843 |
| | | | | 0,6503 |
| | | | | 2,1909 |
| VAR($\epsilon$) | | 1,259 | 0,236 | 0,000 |
| PRE – Maß | | | 81,3% | 100,0% |
| N | | 64 | 64 | 64 |
| D | | 31 | 31 | 31 |
| in % | | 48,4% | 48,4% | 48,4% |
| ln L | | −139,62 | −129,19 | −126,69 |
| Pseudo – $R^2$ | | 0,0% | 7,5% | 9,3% |
| L – Ratio | | 0 | 20,85 | 25,84 |
| df | | 0 | 1 | 4 |

a) Parameter (* sign. 5% – Niveau), b) Standardfehler, c) Antilogarithmus

Zur Illustration dieses Modell verwende ich die KFN – Daten. Die Schätzergebnisse zeigt Tabelle 6.1 (PRG0601). Alle Problemvariablen (mit Ausnahme des Effektes der Schulden) erhöhen das Rückfallrisiko. Allerdings ist nur der Effekt der sozialen Probleme signifikant von Null verschieden. Modell 1, das nur diese Kovariate berücksichtigt, kann daher als ausrei-

chende Beschreibung der Daten akzeptiert werden. Durch Berücksichtigung dieser Variablen läßt sich die Prognose des Bewährungserfolges um 81,3% verbessern. Dazu vergleicht man die geschätzte Fehlervarianz des Basismodells 0 mit der Fehlervarianz in Modell 1, in dem der Einfluß sozialer Probleme kontrolliert wird. Mit den beiden Fehlervarianzen kann man dann in der üblichen Weise ein Maß proportionaler Fehlerreduktion (PRE) berechnen:

$$\text{PRE} = \frac{\hat{\text{VAR}}(\epsilon_0) - \hat{\text{VAR}}(\epsilon_1)}{\hat{\text{VAR}}(\epsilon_0)} = \frac{1{,}259 - 0{,}236}{1{,}259} = 0{,}813$$

LANCASTER (1990: Kap. 4) diskutiert die Probleme von *Mischverteilungsmodellen* im Kontext von Verlaufsanalysen. *Modelle mit unbekannter Heterogenität*, in denen die Fehler seriell miteinander korrelieren und verschiedene Fehlerverteilungen zugelassen sind, findet man bei HECKMAN/SINGER (1984).

## *6.3 Regressionsmodelle für wiederholbare Ereignisse*

Die Likelihood–Funktion (5.46) kann unter bestimmten Annahmen auf wiederholbare Ereignisse verallgemeinert werden. Angenommen eine Person i hat 2 Tätigkeitswechsel zu den Zeitpunkten $t_i^1 = 2$ und $t_i^2 = 8$ (gemessen in Jahren nach Berufseintritt). Bis zum Ende der Untersuchungsperiode $\tau = 10$ konnte kein weiterer Wechsel beobachtet werden. Genauso gut könnte man aber auch die Zeit in Jahren seit Tätigkeitsbeginn (Tätigkeitsdauer) messen: $t_i^1 = 2$, $t_i^2 = 6$, $t_i^3 = 2$. Mit anderen Worten, die Uhr würde mit jedem Wechsel zurückgestellt. Bei wiederholbaren Ereignissen gibt es also mehrere Definitionen der Prozeßzeit. Üblicherweise wird die Zustandsdauer verwendet.

Die Art des Tätigkeitswechsels soll zunächst nicht interessieren (Prozeß mit singulären aber wiederholbaren Ereignissen). Insgesamt hat die Person $E_i = 3$ Episoden. Dementsprechend besteht die Likelihood für diese Folge von Ereignissen und Zensierung aus drei Teilen:

1. aus der Dichte, zum Zeitpunkt $t_i^1 = 2$ ein Ereignis zu haben,
2. aus der Dichte, zum Zeitpunkt $t_i^2 = 6$ ein Ereignis zu haben und schließlich

3. aus der Überlebenswahrscheinlichkeit, bis zum Ende der Untersuchungsperiode keinen weiteren Wechsel zu haben.

Alle drei Funktionen sind wiederum von den Übergangsraten und (das ist jetzt neu) der Vorgeschichte des jeweiligen Ereignisses/Zensierung abhängig. Für die Übergangsraten $r^1(t)$, $r^2(t)$ und $r^3(t)$ läßt sich jeweils ein Regressionsmodell formulieren, das je nach Nummer $e=1,2,3$ der Episode unterschiedliche Kovariaten $\mathbf{x}_i^e$ und Effekte $\boldsymbol{\beta}^e$ enthalten kann. Zusätzlich kann man in dem Regressionsmodell die Geschichte $G^{e-1}$ vor der Episode Nummer e berücksichtigen (z.B. die Anzahl früherer Ereignisse). Über die Übergangsraten werden damit Dichte und Überlebenswahrscheinlichkeit eine Funktion von $x_i^e$, $\beta^e$ und $G_i^{e-1}$. Der individuelle Beitrag der beschriebenen Person i berechnet sich dann wie folgt:

$$L_i = f(t_i^1=2|x_i^1,\beta^1) \cdot f(t_i^2=6|x_i^2,\beta^2,G_i^1) \cdot (S(t_i^3=2|x_i^3,\beta^3,G_i^2)$$

Wenn diese zwei Ereignisse und eine Zensierung unabhängig voneinander sind oder man die serielle Abhängigkeit in einem Regressionsmodell kontrolliert, dann kann man diese zusammenhängende Ereignisgeschichte wie drei voneinander unabhängige Beobachtungen behandeln: D.h. drei verschiedene Personen, von denen die erste die Tätigkeit nach 2 Jahren wechselt, die zweite nach 6 Jahren und die dritte nach 2 Jahren immer noch dieselbe Tätigkeit ausübt. Wenn diese Annahme richtig ist, dann sind alle bisherigen Likelihood-Funktionen anwendbar, mit dem einzigen Unterschied, daß jetzt jede Episode einer Person wie eine neue Untersuchungseinheit behandelt wird. Eine entsprechende Analysedatei hätte dann genauso viele Fälle wie insgesamt Tätigkeiten auftreten. Für Prozesse mit multiplen Ereignissen, deren Likelihood-Funktion in (5.46) definiert ist, müßte also nur das Produkt über den Index i an die größere Fallzahl angepaßt werden. Formal exakt ließe sich dieser Tatbestand durch ein weiteres Produkt über alle $e=1,\ldots,E_i$ Tätigkeiten pro Person berücksichtigen:

$$(6.6) \qquad L = \prod_j^J \prod_k^K \prod_i^N \prod_e^{E_i} S_{jk}(t_i^e|\mathbf{x}_{ijk}^e,\boldsymbol{\beta}_{jk}^e,G_i^{e-1})\, r_{jk}^e(t_i^e|\mathbf{x}_{ijk}^e,\boldsymbol{\beta}_{jk}^e,G_i^{e-1})^{\delta_{ijk}^e}$$

$$\text{mit } \delta_{ijk}^e = \begin{cases} 1 & \text{(Wechsel von j nach k in Episode e)} \\ 0 & \text{(Zensierung)} \end{cases}$$

Auf diese Weise erhält man eine *allgemeine Likelihood-Funktion für multiple und wiederholbare Ereignisse*. Die obige Unabhängigkeitsannahme ist natürlich eine sehr schwerwiegende Voraussetzung. Sie ist eigentlich nur dann gegeben, wenn man alle (beobachteten und unbeobachteten) Merkmale der Personen und die Vorgeschichte kontrollieren kann. HAMERLE (1989) diskutiert Regressionsmodelle mit wiederholbaren Ereignissen und zeigt, wie sie mit gängigen Programmpaketen geschätzt werden können.

## 6.4 *Regressionsmodelle für zeitdiskrete Verlaufsdaten*

*Regressionsmodelle für diskrete Zeitdauern* können mit WLS- oder ML-Schätzungen überprüft werden. Die Berechnungen in Kapitel 4 verweisen im Prinzip schon auf den ersten der beiden Ansätze: Ergebnisse von Sterbetafelschätzungen werden dabei als abhängiges Merkmal in ganz normalen Regressionsanalysen verwendet. Dabei ist jedoch zu berücksichtigen, daß das abhängige Merkmal $\hat{S}(t)$ (ähnlich wie ein Anteilswert) nicht die schönen Eigenschaften der Streuungsgleichheit besitzt, wie die üblichen metrischen abhängigen Merkmale. Daher sind entsprechend verallgemeinerte Kleinste-Quadrate-Schätzungen wie z.B. eine gewichtete Regression (WLS - weighted least squares) notwendig (vgl. GEHAN/SIDDIQUI 1973).

Die Ausgangsdaten einer Sterbetafel, insbesondere wenn man nach verschiedenen Subgruppen differenziert, lassen sich auch als multivariate Kreuztabelle interpretieren (vgl. Tabelle 3.8). Die Sterbetafelschätzer für S(t) etc. sind daher nichts anderes als abgeleitete Funktionen der Häufigkeiten der multivariaten Kreuztabelle. Das ermöglicht den Einsatz von *Techniken multivariater Kreuztabellenanalyse*. Ein Ansatz, der von GRIZZLE, STARMER und KOCH (1969) entwickelt wurde und nach den Autoren als GSK-Ansatz bezeichnet wird, gestattet auf sehr flexible Art und Weise, komplizierte Funktionen der Häufigkeiten einer multivariaten Kreuztabelle zu untersuchen. Dabei verwendet man eine WLS-Schätzung. Die Anwendung des GSK-Ansatzes auf Sterbetafeln diskutieren KOCH/JOHNSON/TOLLEY (1972) und JOHNSON/KOCH (1978). Eine Anwendung auf Daten zur Arbeitslosigkeitsdauer findet sich bei ANDRESS (1985).

Der zweite Ansatz geht davon aus, daß die Daten im Prinzip aus zweierlei Informationen bestehen: a) diskrete Zeitdauern ohne Ereignis, b) diskrete Zeitdauern mit Ereignis. Bei singulären Ereignissen entspricht

dieses Vorgehen dem üblichen *logistischen Regressionsmodell*. Bei multiplen Ereignissen ergibt sich das multinomiale logistische Regressionsmodell. Die gesuchten Parameter dieser Modelle können dann mit ML-Schätzungen berechnet werden.

Dieser Ansatz hat zunächst nicht die Beschränkung auf ausschließlich diskrete Merkmale. Hier sind auch metrische Merkmale bei den unabhängigen Variablen zugelassen. Sollte es sich dennoch um ausschließlich nichtmetrische Merkmale handeln, dann ergeben sich die bekannten ML-Verfahren multivariater Kreuztabellenanalyse, die in der Soziologie vor allem durch die Arbeiten von GOODMAN (1978) bekannt wurden (*log-lineare Modelle*). Eine Einführung in diesen zweiten Ansatz findet sich bei ALLISON (1982) und HAMERLE/TUTZ (1989).

## 6.5 *Residuenanalyse*

Eine Möglichkeit, Residuen im Rahmen von Verlaufsanalysen zu definieren, bietet die *kumulierte Rate* H(t). Sie kommt in Gleichung (2.13) vor. Abstrahieren wir einmal vom Kontext dieser Gleichung und ersetzen wir darin H(t) durch $\epsilon$:

$$S(\epsilon) = \exp(-1 \cdot \epsilon) \quad \text{mit } \epsilon = H(t) \tag{6.7}$$

So gesehen, sieht (6.7) aus wie eine Exponentialverteilung mit $\lambda = 1$. Anders ausgedrückt, die Zufallsvariable $\epsilon = H(t)$ folgt einer Standardexponentialverteilung mit Parameter $\lambda = 1$.

Dieses Ergebnis wird nun bei der folgenden Residuenanalyse berücksichtigt. Man sagt einfach (definiert), $H(t_i)$ sei das *Residuum* $\epsilon_i$ der Beobachtung i. Wenn die vorhergehenden Überlegungen richtig sind, dann müssen diese Residuen standard-exponentialverteilt sein, ähnlich wie die standardisierten Residuen bei klassischen Regressionsmodellen standard-normalverteilt sind. Diese Annahme kann man dann mit Hilfe graphischer Verfahren testen.

Der Vorteil dieser "Residuen" ist, daß die kumulierte Rate $H(t_i)$ für jede einzelne Beobachtung i berechnet werden kann, während der oben beschriebene Vergleich von Modellprognosen mit nicht-parametrischen Schätzungen immer eine Gruppierung des Datenmaterials voraussetzt. Prinzipiell ist also eine differenziertere Analyse auf der Ebene einzelner Beob-

achtungen möglich. Dabei stellt sich aber das Problem, wie zensierte Beobachtungen behandelt werden sollen. Eine Vereinbarung lautet, daß Residuen, die zu einer zensierten Beobachtung gehören, als zensierte Residuen betrachtet werden sollen. Möchte man daher bestimmte Verteilungsannahmen über die Residuen testen, dann muß man die in Kapitel 4 diskutierten nicht-parametrischen Verfahren für zensierte Daten verwenden (Kaplan-Meier-Schätzer, empir. kumul. Risikofunktion). Man berechnet dazu auf Grund der Schätzergebnisse die Residuen $\hat{e} = \hat{H}(t_i)$ und zeichnet die empir. kumulierte Risikofunktion dieser Residuen. Wenn die obige Verteilungsannahme (Standardexponentialverteilung) richtig ist, dann muß diese Funktion einer Geraden mit Steigung 1 entsprechen (vgl. die Beispiele bei BLOSSFELD et al. 1986). In einigen Programmen kann man diese und andere Residuen direkt abrufen. Weitere Hinweise zur Residuenanalyse finden sich bei LAWLESS (1982: 281f.). Kritisch äußern sich LAGAKOS (1980) und CROWLEY/STORER (1983).

Anstelle einer Residuenanalyse kann man auch direkt einen Fehlerterm in der Regressionsgleichung (5.2d) berücksichtigen, der alle nicht berücksichtigten Einflüsse und Meßfehler erfaßt (*unbekannte Heterogenität*). Auf diese Weise könnte man anlog dem Determinationskoeffizienten R-Quadrat die Reduktion der Fehlervarianz berechnen, die sich mit der Kontrolle bekannter Heterogenität, d.h. mit Kontrolle der Kovariaten ergibt. Ergebnis wäre ein PRE-Maß über den *Modellfit*. Ein ganz einfaches Modell dieses Typs wurde in Abschnitt 6.2 vorgestellt.

# *Anhang*

## *Anhang A: Mathematischer Anhang*

Einige Ableitungen im Text setzen Kenntnisse der Differential–, Integral–und Wahrscheinlichkeitsrechnung voraus. An dieser Stelle sollen noch einmal die wesentlichen mathematischen Grundbegriffe wiederholt werden. Dieser Anhang kann natürlich die entsprechende Fachliteratur nicht ersetzen. Einzelne Gleichungen des Textes sollen lediglich von einem mathematischen Standpunkt motiviert werden. Vielleicht erinnert sich ja auch der eine oder andere Leser an früher erworbene Kenntnisse der Schulmathematik.

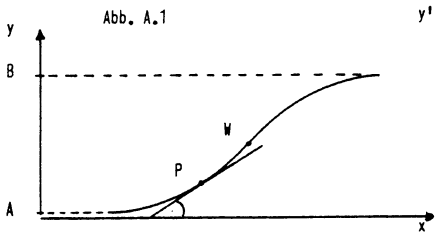
### *Differentialrechnung*

Zwei Beispiele sollen die Grundfragestellung der Differentialrechnung (DR) verdeutlichen.

Beispiel 1: Eine Straßenbahn fährt von Haltestelle A nach Haltestelle B. Sie fährt langsam an, beschleunigt bis sie in der Mitte zwischen beiden Haltestellen die volle Geschwindigkeit erreicht, verringert ab da ihr Tempo und fährt schließlich langsam in die zweite Haltestelle ein. Das zugehörige Weg–Zeit–Diagramm findet sich in Abbildung A.1 (y=Weg, x=Zeit). Die momentane Geschwindigkeit der Straßenbahn hängt offenbar mit der Steilheit der Kurve zusammen. Diese ergibt sich durch das Steigungsmaß einer Tangenten, die in jedem beliebigen Punkt P an die Kurve gelegt werden kann. In Abbildung A.2 ist dieses Steigungsmaß y' für jeden Wert x abgetragen. Man erkennt, wie die Geschwindigkeit bis zum Punkt W zunimmt und danach wieder abnimmt.

Beispiel 2: In Abbildung A.1 könnte auch der kumulierte Anteil der Personen mit mindestens einem Tätigkeitswechsel dargestellt sein (y=kumulierter Anteil, x=Zeit). Zu Beginn des Prozesses A hat keine Person ihre Tätigkeit gewechselt. Am Ende B hat jede Person mindestens einen Wechsel erlebt. Die Abbildung beschreibt die Veränderung des Anteils im Zeitablauf. Zu Beginn ändert er sich kaum, dann sehr schnell und je mehr er sich der oberen Grenze B nähert, um so geringer sind wiederum die Änderungen. Aus diesem Verlauf des Anteilswertes kann man folgern, daß die Wahrscheinlichkeit eines Wechsels in der Mitte des Intervalls (um den

Punkt W) am größten ist. Die zweite Abbildung A.2 zeigt die Veränderung des kumulierten Anteils y.



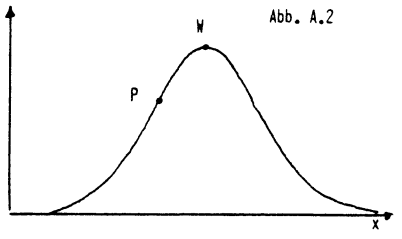



Abb. A.2

Die Beispiele demonstrieren Anwendungen der Differentialrechnung in der Physik (Geschwindigkeitsmessung) und in der Statistik (Wahrscheinlichkeitsrechnung). Weitere Anwendungen findet man in der numerischen Optimierung (Bestimmung der Extrema von Funktionen, s.unten). Das mathematische Ausgangsproblem ist die Frage, wie man die Steigung einer Tangenten in einem gegebenen Punkt $P_0$ einer beliebigen Kurve bestimmt.

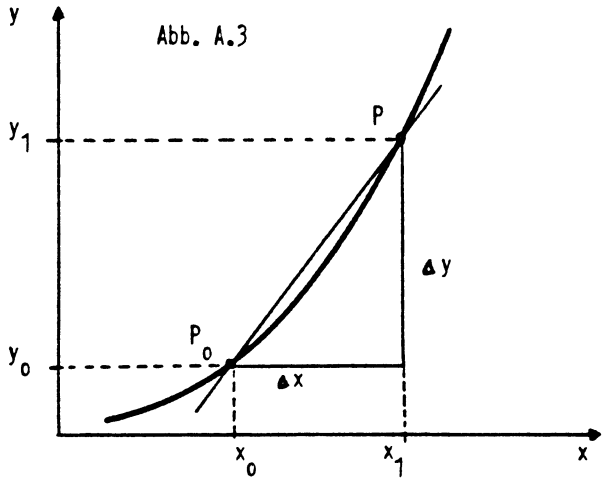
Eine Kurve kann als das Bild einer Funktion $y = f(x)$ betrachtet werden. Eine Gerade durch die Punkte $P_0$ und P schneidet diese Kurve (vgl. Abbildung A.3). Die Steigung dieser Geraden entspricht dem Quotienten $\Delta y/\Delta x$. Man bezeichnet ihn als sogenannten *Differenzenquotient*:

$$\text{(A.1)} \qquad \frac{\Delta y}{\Delta x} = \frac{y_1 - y_0}{x_1 - x_0} = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

Je näher jetzt der Punkt P an $P_0$ heranrückt, um so mehr entspricht die Gerade der gesuchten Tangenten im Punkt $P_0$. Der Differenzenquotient drückt also das Verhalten der Kurve im Punkt $P_0$ um so besser aus, je näher P und $P_0$ zusammenliegen. $\Delta x$ wird dabei immer kleiner. Man kann sich überlegen, daß $\Delta x$ ab einem bestimmten Punkt gleich Null ist. Durch diese Grenzbetrachtung erhält man schließlich die 1. Ableitung der Funktion $y = f(x)$ an der Stelle $x_0$:

$$\text{(A.2)} \qquad f'(x_0) = \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

Genauer gesagt muß man zeigen, daß der Grenzwert bei rechts– und linksseitiger Annäherung gleich ist. Der Differenzenquotient wird zum sogenannten *Differentialquotienten.* Er mißt umgangssprachlich die momentane Änderung einer variablen Größe.




Existiert der Grenzwert (A.2) nicht für einen Punkt sondern für einen bestimmten Wertebereich, dann ist die Funktion in diesem Bereich differenzierbar. Für jeden Wert x kann der Wert des Differentialquotienten angegeben werden, der somit selbst eine Funktion von x ist. Man nennt daher diese abgeleitete Funktion die *Ableitung* oder den Differentialquotienten. Hierfür haben sich mehrere Schreibweisen eingebürgert:

$$y' = f'(x) = \frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} \tag{A.3}$$

Auch diese Funktion läßt sich erneut ableiten. Man erhält die 2. Ableitung f"(x) usw. dy und dx bezeichnet man auch als *Differentiale.*

Damit eine Funktion differenzierbar ist, müssen bestimmte mathematische Voraussetzungen gegeben sein. *Stetigkeit* ist z.B. eine notwendige, aber nicht hinreichende Bedingung. Für die wichtigsten elementaren Ableitungen gibt es Nachschlagewerke. Einige der folgenden Beispiele werden von mir häufiger benutzt:

(A.4) $y = a + by \rightarrow y' = b$

$y = \exp x \rightarrow y' = \exp x$

$y = \ln x \rightarrow y' = \frac{1}{x}$

Für komplexere zusammengesetzte Funktionen gibt es verschiedene Differenzierungsregeln (Faktor−, Summen−, Produkt−, Quotienten−, Potenz−, Umkehr− und Kettenregel). Z.B. zeigt (A.5) nicht den Logarithmus von x sondern den Logarithmus einer Funktion von x:

(A.5) $y = \ln (a + b \exp x)$

Hier hilft die sogenannte *Kettenregel*:

(A.6) $y = f(z) \quad \text{mit } z = g(x) \rightarrow y' = f'(z) \cdot g'(x)$

Die Ableitung einer solchen zusammengesetzten Funktion ist das Produkt der "inneren" und der "äußeren" Ableitung:

(A.7) $y = \ln z \quad \text{mit } z = a + b \exp x \rightarrow y' = \frac{b \exp x}{a + b \exp x}$

$$\text{wegen } y' = \frac{1}{z} \text{ und } z' = b \exp x$$

Die DR dürfte den meisten noch unter dem Stichwort "Kurvendiskussion" aus der Schule bekannt sein (Bestimmung der Extrema und Wendepunkte). Damit ein Minimum vorliegt, muß die 1. Ableitung einer Funktion gleich Null und die 2. Ableitung größer als Null sein. LS−Schätzungen sind z.B. eine Anwendung dieses Prinzips. Hier muß das Minimum der quadrierten Residuen gefunden werden. Dabei ist das Problem, daß die abhängige Variable y eine Funktion mehrerer unabhängiger Variablen ist (multivariate Zusammenhänge). Man muß also ein Minimum finden, das für alle Parameter gleichermaßen gilt. Hierzu verwendet man das Hilfsmittel der *partiellen Differentiation*.

Als Beispiel verwende ich eine einfache linear−additive Funktion mit zwei unabhängigen Variablen:

(A.8) $y = a + bx + cz$

Diese Funktion kann nach jeder unabhängigen Variablen abgeleitet werden, wobei die anderen unabhängigen Variablen jeweils wie Konstanten behandelt werden. Die 1. partielle Ableitung nach x lautet danach (z konstant):

(A.9a) $$\frac{\partial y}{\partial x} = b$$

und die 1. partielle Ableitung nach z ist dementsprechend (x konstant):

(A.9b) $$\frac{\partial y}{\partial z} = c$$

*Integralrechnung*

Um die grundlegende Fragestellung der Integralrechnung (IR) zu verdeutlichen, betrachte ich eine Umkehrung der obigen Beispiele.

Beispiel 1: Angenommen man möchte von der momentanen Geschwindigkeit einer Straßenbahn auf den zurückgelegten Weg schließen. Abbildung A.4 zeigt dazu noch einmal das obige Geschwindigkeitsdiagramm in etwas anderer Form (y = Geschwindigkeit, x = Zeit). Der Rückschluß von der Geschwindigkeit auf den zurückgelegten Weg ist nämlich dann besonders gut vorstellbar, wenn die Geschwindigkeit nicht kontinuierlich regelbar ist, sondern nur in Stufen geschaltet werden kann (man erinnere sich an die Stufenschalter in den alten Straßenbahnen). Den zurückgelegten Weg erhält man dann, indem man das Zeitintervall $\Delta x_i$, in dem eine bestimmte Geschwindigkeitsstufe gefahren wurde, mit der entsprechenden Geschwindigkeit $y = f(x_i)$ multipliziert. Dieses Produkt entspricht genau der Fläche des schraffierten Rechteckes. Der Gesamtweg ergibt sich schließlich, indem man die Flächen aller Rechtecke zusammenaddiert. Der zurückgelegte Weg, der sich auf Grund dieser Berechnung für jeden Zeitpunkt ergibt, ist in Abbildung A.5 dargestellt.

Beispiel 2: Aus der Veränderung eines (kumulierten) Anteilswertes (z.B. Anteil der Personen mit mindestens einem Tätigkeitswechsel) kann man sowohl die Wahrscheinlichkeit bestimmter Ereignisse berechnen, als auch die Wahrscheinlichkeit, bis zum Zeitpunkt t mindestens ein Ereignis zu erleben. Abbildung A.4 ist jetzt quasi ein Stäbchendiagramm, in dem die relativen Häufigkeiten von Tätigkeitswechseln pro Zeiteinheit abgetragen

sind (y=relative Häufigkeit, x=Zeit). Die Wahrscheinlichkeit, bis zum Zeitpunkt T=t mindestens ein Ereignis zu erleben, ist wiederum die Summe der relativen Häufigkeiten vom Zeitpunkt 0 bis zum Zeitpunkt t. Abbildung A.5 entspricht also dem Verlauf dieser kumulierten Wahrscheinlichkeit.

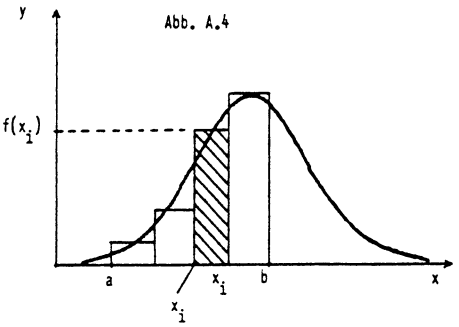


Abb. A.4

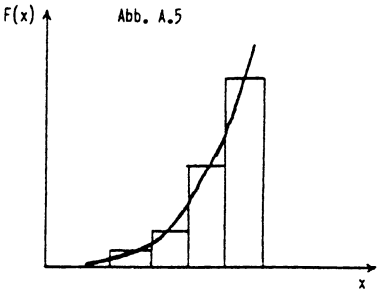


Abb. A.5

Das allgemeine Thema der IR ist also die Bestimmung der Fläche unter einer Kurve. Auch hier finden sich die verschiedensten Anwendungen in der Physik und Statistik, wie die Beispiele zeigen. In gewisser Weise ist die IR die Umkehrung der DR. Sehr oberflächlich gesprochen geht es im einen Fall um die Berechnung von Differenzen (DR) und im anderen Fall um die Berechnung von Summen (IR).

Vom mathematischen Standpunkt ist das Tangentenproblem die Ausgangsfrage der DR. Dagegen beschäftigt sich die IR mit dem Quadraturproblem (Bestimmung von Flächeninhalten). Wie in Abbildung A.4 angedeutet, löst man diese Frage durch Annäherung der kontinuierlichen Funktion durch eine Treppenfunktion. Genauer gesagt werden die Rechtecke so gezeichnet, daß sie einmal genau unter die Kurve passen (Untersumme) und das andere Mal gerade darüber (Obersumme). Die Fläche jedes Rechtecks ist wie gesagt $f(x_i)\cdot\Delta x_i$ und die Gesamtfläche innerhalb eines Intervalls von a nach b ergibt sich aus der Summe aller Rechtecke im Intervall. Je mehr Rechtecke man verwendet, desto kleiner wird $\Delta x_i$ und desto besser wird der Flächeninhalt der kontinuierlichen Kurve angenähert. Wenn sich bei

dieser Grenzbetrachtung sowohl bei der Annäherung von oben (Obersumme) als auch von unten (Untersumme) der gleiche Grenzwert ergibt, dann ist die Funktion im Intervall [a,b] integrierbar:

$$\lim_{n\to\infty} \sum_{i=1}^{n} f(x_i)\Delta x_i = \int_a^b f(x)dx \tag{A.10}$$

Dieser Grenzwert wird als das *bestimmte Integral* der Funktion f(x) im Intervall [a,b] bezeichnet. Die Funktion f(x) wird *Integrand*, die Größe x *Integrationsvariable* genannt. a und b sind die *Intervallgrenzen*, innerhalb derer die Fläche (das Integral) berechnet wird (daher bestimmtes Integral). Im Text ergeben sich manchmal die gleichen Symbole für die Intervallgrenzen und die Integrationsvariable (z.B. Integral von $f(t)=\exp(-\lambda t)$ im Intervall 0 bis t). Ich verwende daher u als Integrationsvariable.

Jede stetige Funktion ist integrierbar und für das Rechnen mit bestimmten Integralen gelten wiederum bestimmte Regeln (z.B. für die Summe von Integralen, für die Multiplikation mit einem konstanten Faktor oder für die Vertauschung der Intervallgrenzen). Um das Integral einer Funktion f(x) zu bestimmen, benutzt man die Tatsache, daß die Integration die Umkehrung der Differentiation ist: Ist nämlich eine Funktion F(x) im betrachteten Intervall stetig und differenzierbar, dann entspricht ihre Ableitung dem Wert des Integranden an der oberen Grenze.

$$F(x) = \int_a^x f(u)du \tag{A.11}$$

$$F'(x) = \frac{d}{dx}\int_a^x f(u)du = f(x)$$

Diese Funktion F(x), deren Differentialquotient der Integrand ist, wird *Stammfunktion* genannt. Die Aufgabe reduziert sich also auf die Frage, welche Funktion F(x) zu f(x) paßt, so daß ihre 1. Ableitung f(x) entspricht.

Man beachte aber, daß mehrere Funktionen, die sich nur durch eine Konstante unterscheiden, die gleiche Ableitung haben. Z.B.:

(A.12) $y = 3x + 4 \quad \rightarrow \quad y' = 3$
$y = 3x + 100 \quad \rightarrow \quad y' = 3$

Daher gehört zu jeder stetigen Funktion f(x) eine ganze Schar von Stammfunktionen, die sich nur durch eine Konstante $\varkappa$ unterscheiden:

(A.13) $\int a\,dx = ax + \varkappa$
$\int \exp x\,dx = \exp x + \varkappa$
$\int x^{-1}\,dx = \ln x + \varkappa$

In dieser Form gibt es Nachschlagewerke, wo man für die verschiedensten Funktionen f(x) die entsprechende Stammfunktion F(x) finden kann. Dabei wird die Konstante $\varkappa$ der Einfachheit halber weggelassen.

An dieser Stelle begnüge ich mich mit den elementaren Stammfunktionen, die im Text häufiger vorkommen. Für komplexere Funktionen gibt es wiederum bestimmte Integrationsverfahren (partielle Integration, Substitutionsmethode). Die Integration einer Funktion ist in der Regel schwieriger lösbar als ihre Differentiation. Ich verzichte daher auf eine Erwähnung einzelner Integrationsverfahren. Der Leser kann jedoch die Integrale im Text jederzeit überprüfen, indem er die Probe durch erneute Differentiation macht. Es muß sich wieder die Ausgangsfunktion ergeben.

Bei den Integralen in (A.13) wurden die Intervallgrenzen fortgelassen. Man bezeichnet sie daher als *unbestimmte Integrale*. Die entsprechenden bestimmten Integrale erhält man, indem man die Intervallgrenzen in die Stammfunktion einsetzt und den Wert der Stammfunktion an der unteren Grenze von dem an der oberen Grenze abzieht:

(A.14) $\int_1^5 a\,dx = |ax + \varkappa|_1^5 = (5a + \varkappa) - (1a + \varkappa) = 4a$

Bei dieser Rechnung fällt die Konstante $\varkappa$ übrigens weg.

*Differentialgleichungen*

Besteht zwischen einer Funktion einer oder mehrerer Variablen und ihren Ableitungen eine Beziehung in Form einer Gleichung, in der auch die

Variablen selbst vorkommen können, dann spricht man von einer *Differentialgleichung* (DGL). Man unterscheidet DGL danach,

- ob nur eine oder mehrere Variablen auftreten (gewöhnliche versus partielle DGL).
- welchen Grad die einzelnen Ableitungen haben (Ordnung der DGL).
- ob die Ableitungen miteinander multipliziert oder addiert werden (lineare versus nicht-lineare DGL).
- ob die DGL nach einer Ableitung aufgelöst ist (explizite versus implizite DGL).

Danach treten im Text nur gewöhnliche, explizite und lineare DGL 1. Ordnung auf (vgl. Gleichungen 2.12 und 5.51).

Genauer gesagt handelt es sich um homogene DGL, weil ein von der Funktion und deren 1. Ableitung unabhängiger Teil fehlt. Gleichung (2.12) ist dafür ein gutes Beispiel:

$$\frac{d\,S(t)}{dt} = -r(t)\,S(t) \tag{A.15}$$

Es geht um die Überlebenswahrscheinlichkeit S(t), die eine Funktion der Variablen t (Zeit) ist. Die Gleichung ist nach der Ableitung dieser Funktion aufgelöst (explizite DGL). Außerdem tritt noch eine weitere Funktion r(t) der Variablen t auf, die jedoch nicht von S(t) unabhängig ist, sondern mit S(t) multipliziert wird (homogene DGL).

Diese Art von DGL läßt sich relativ einfach mit folgendem Verfahren lösen. Dazu isoliert man die beiden Differentiale dS(t) und dt auf jeweils einer Seite der Gleichung:

$$\frac{d\,S(t)}{S(t)} = -r(t)\,dt \tag{A.16}$$

Durch Umkehrung der Differentiation (Integration beider Seiten der Gleichung) kann man dann die beiden Differentiale auflösen:

$$\int_{S(0)}^{S(t)} S(u)^{-1}\,dS(u) = \int_{0}^{t} -r(u)\,du \tag{A.17}$$

Lediglich bei den Intervallgrenzen muß man etwas aufpassen. Grundsätzlich verwendet man die untere Grenze bzw. den aktuellen Wert der betrachteten Variablen. In diesem Fall ist das die Zeit t, die nur positive Werte annehmen kann und daher den Wert 0 als untere bzw. t als obere Grenze hat. Die Variable selbst ist jedoch nicht immer die Integrationsvariable. Häufig ist es eine Funktion derselben (wie S(u) in (A.17) auf der linken Seite). Man muß daher die entsprechenden Funktionswerte einsetzen, also S(0) und S(t).

Die beiden Integrale in Gleichung (A.17) sind leicht gelöst. Unter Berücksichtigung der Konstanten $\varkappa$, die ich auf der rechten Seite der Gleichung zusammenfasse, ergibt sich:

$$\ln S(t) = -\int_0^t r(u)\,du + \varkappa \tag{A.18}$$

Verbleibt nur die Frage, wie man die Integrationskonstante $\varkappa$ bestimmt (Anfangswertproblem). Dazu ist es hilfreich, die Anfangsbedingungen des Prozesses zu untersuchen. Da es sich bei S(t) um eine Überlebenswahrscheinlichkeit handelt, muß diese Größe zu Beginn ($t=0$) gleich 1 sein. Durch Einsetzen von $S(0)=1$ ergibt sich dann:

$$\ln S(0) = -\int_0^0 r(u)\,du + \varkappa \quad \rightarrow \quad 0 = 0 + \varkappa \tag{A.19}$$

Folglich ist die Integrationskonstante unter diesen Anfangsbedingungen Null. Die Lösung der DGL (5.51) verläuft in ähnlicher Weise, jedoch ist der Anfangswert unbekannt, so daß die Konstante $\varkappa$ erhalten bleibt.

*Exponentialfunktion und natürlicher Logarithmus*

Die Exponentialfunktion und der natürliche Logarithmus treten an relativ vielen Stellen im Text auf, so daß sich die Frage aufdrängt, ob es dafür besondere Gründe gibt. Wie man aus Gleichung (A.4) erkennt, ist die Ableitung der Exponentialfunktion gleich der Funktion selbst: $y=y'=\exp(x)$. Folglich führen alle Naturvorgänge, bei denen die Veränderung einer Größe gleich oder proportional ihrem absoluten Wert ist, auf diese Funktion. Angenommen das Bevölkerungswachstum dN/dt sei dem

Bevölkerungsbestand N proportional, dann ergibt sich folgende DGL ($\gamma$ sei ein Proportionalitätsfaktor):

$$\text{(A.20)} \qquad \frac{dN}{dt} = \gamma N \quad \rightarrow \quad N = \exp(\gamma t) + x$$

Ihre Lösung ist wiederum die Exponentialfunktion. Ähnliche Argumente gelten für viele andere soziale Prozesse und Naturvorgänge (z.B. radioaktiver Zerfall).

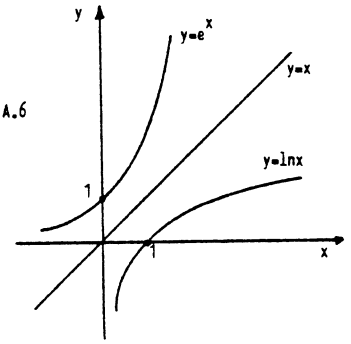


Abb. A.6

Die Umkehrung der Exponentialfunktion ist der natürliche Logarithmus. Es gilt daher $\ln(\exp x) = x$ bzw. $\exp(\ln x) = x$. Beide Funktionen sind in Abbildung A.6 dargestellt. Durch Spiegelung an der Diagonalen $y = x$ ergibt sich eine Funktion aus der anderen. Die *Exponentialfunktion* ist ursprünglich als unendliche Reihe definiert:

$$\text{(A.21)} \qquad e^x = \exp x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots$$

Für $x = 1$ ergibt sich die Konstante $e = 2{,}718\ldots$ Es gilt auch:

(A.22) $$e^x = \exp x = \lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n$$

Diese Reihe hat gewisse Ähnlichkeiten mit der Zinseszins-Formel. Sie wächst jedoch noch schneller.

Der natürliche Logarithmus ergibt sich schließlich durch Anwendung der Logarithmendefinition:

(A.23) $$\log_b a = c \longleftrightarrow a = b^c$$

Als Logarithmus c der Zahl a zur Basis b wird der Exponent der Potenz bezeichnet, in die man b erheben muß, um die Zahl a zu erhalten. Verwendet man als Basis die Konstante e, dann erhält man den *natürlichen Logarithmus*:

(A.24) $$\ln a = c \longleftrightarrow a = e^c = \exp c$$

Auch der natürliche Logarithmus läßt sich als unendliche Reihe darstellen. Im Text verwende ich folgende Formel, die allerdings nur für x-Werte zwischen $-1$ und 1 gilt:

(A.25) $$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots -$$

Exponentialfunktion und natürliche Logarithmen sind also nichts anderes als spezielle Potenzen und Logarithmen. Man kann daher mit ihnen rechnen wie mit normalen Potenzen und Logarithmen. Der Vollständigkeit halber habe ich noch einmal die wichtigsten Rechenregeln zusammengestellt:

(A.26)
$$\exp a \cdot \exp b = \exp(a + b)$$
$$\frac{\exp a}{\exp b} = \exp(a - b)$$
$$(\exp a)^b = (\exp b)^a = \exp(ab)$$
$$\exp(-a) = \frac{1}{\exp a}$$
$$\exp \frac{a}{b} = \sqrt[b]{\exp a}$$

$$\ln(ab) = \ln a + \ln b$$

$$\ln \frac{a}{b} = \ln a - \ln b$$

$$\ln(a^b) = b \ln a$$

$$\ln \sqrt[b]{a} = \frac{\ln a}{b}$$

*Zufallsvariablen*

Die Verbindungen der Wahrscheinlichkeitsrechnung zur DR und IR sind schon aus den vorhergehenden Beispielen deutlich geworden. An dieser Stelle möchte ich noch einmal explizit einige Begriffe erläutern.

Man nennt eine Größe zufällig oder *Zufallsvariable*, wenn sie bei verschiedenen, unter gleichen Bedingungen durchgeführten Versuchen verschiedene Werte (*Realisationen*) annehmen kann, von denen dann jeder Wert ein zufälliges Ergebnis ist. Dabei kann eine Zufallsvariable in einem Intervall entweder endlich viele oder beliebig viele Werte annehmen. Dementsprechend unterscheidet man zwischen *diskreten* und *kontinuierlichen* Zufallsvariablen. Beispiele sind die Augenzahl beim Würfeln (diskrete Z.) oder der Intelligenzquotient in einer Gruppe von Personen (kontinuierliche Z.).

Abb. A.7 Wahrscheinlichkeitsfunktion

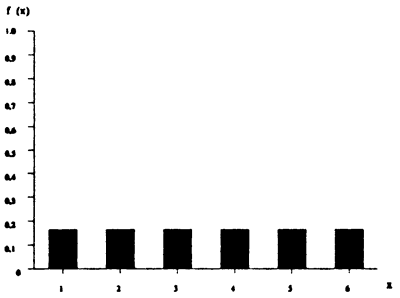
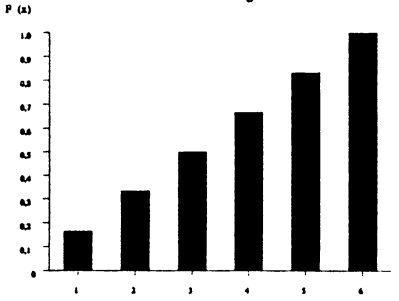


Abb. A.8 Verteilungsfunktion



Eine Zufallsvariable X ist erst dann vollständig charakterisiert, wenn die Wahrscheinlichkeit $P(X=x)$ jeder möglichen Realisation x bekannt ist (zur besseren Unterscheidung wird die Zufallsvariable groß und ihre Realisation

klein geschrieben). Liegen diese Angaben vor, dann ist das *Verteilungsgesetz* oder die *Verteilung* der Zufallsvariablen bekannt. Für einen idealen Würfel ergibt sich z.B. die Verteilung in Abbildung A.7.

Die diskrete Wahrscheinlichkeitsfunktion f(x) beschreibt die Wahrscheinlichkeit jeder Realisation. In diesem Fall ist also jede Augenzahl gleich wahrscheinlich. Die (kumulierte) Verteilungsfunktion mißt die Wahrscheinlichkeit, daß die gewürfelte Augenzahl kleiner oder gleich einer bestimmten Zahl ist (vgl. Abbildung A.8). Sie ergibt sich, indem man die entsprechenden Einzelwahrscheinlichkeiten f(x) von der kleinsten möglichen Realisation bis zur fraglichen Augenzahl summiert. Umgekehrt kann man auch aus der Verteilungsfunktion F(x) die Funktion f(x) errechnen. Die Wahrscheinlichkeit $f(x=2)$ für zwei Augen ergibt sich aus der Differenz der entsprechenden Werte der Verteilungsfunktion $F(x=2)-F(x=1)$.

Entsprechende Überlegungen gelten für kontinuierliche Zufallsvariablen. Angenommen der Intelligenzquotient ist normalverteilt. Die Dichtefunktion f(x) entspricht dann der Normalverteilung (vgl. Abbildung A.9) und die Verteilungsfunktion entspricht der kumulierten Normalverteilung (vgl. Abbildung A.10). Wie in den vorhergehenden Beispielen angedeutet, verwendet man jetzt Integration und Differentiation statt Summen und Differenzen:

$$\text{(A.27)} \qquad F(x) = \int_{-\infty}^{x} f(u)\,du \quad \text{bzw.} \quad f(x) = F'(x)$$

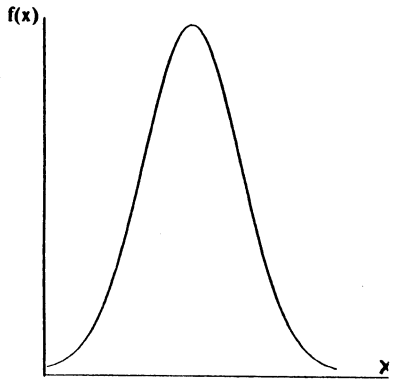
Der Ausdruck f(x)dx entspricht der diskreten Wahrscheinlichkeitsfunktion. Die Wahrscheinlichkeit, daß eine kontinuierliche Zufallsvariable in einem bestimmten Wertebereich ($a \leq x \leq b$) liegt (die Wahrscheinlichkeit einzelner (!) Realisationen ist bei kontinuierlichen Zufallsvariablen Null), ergibt sich dann durch Integration der Dichte f(x) innerhalb der Grenzen a und b. Die Dichte f(x) hat nur positive Werte und das Integral über den gesamten Wertebereich ergibt 1. Daher gilt:

$$\text{(A.28)} \qquad f(x) \geq 0 \qquad \text{und} \qquad \int_{-\infty}^{\infty} f(x)\,dx = 1$$

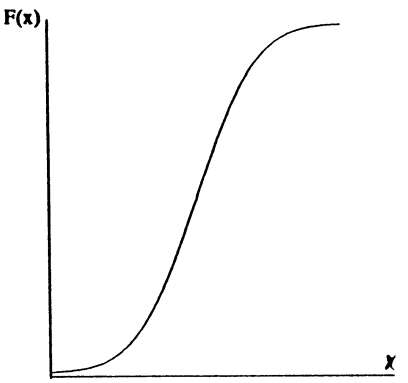
Daher hat die Verteilungsfunktion F(x) bei minus unendlich den Wert 0 und bei plus unendlich den Wert 1 (mit dieser Anfangsbedingung entfällt

die Konstante $x$ in dem Integral A.27). Dazwischen nimmt F(x) monoton zu.

A.9 Dichte




A.10 Verteilungsfunktion




Zur Charakterisierung einer Zufallsvariablen eignet sich auch der *Erwartungswert*, d.h. der Wert, den man bei Berücksichtigung aller möglichen Realisationen und ihrer Wahrscheinlichkeiten im Mittel erwartet. Man berechnet ihn analog einem gewichteten Durchschnitt, jedoch dienen jetzt die Wahrscheinlichkeiten als Gewichte. Für das Würfelbeispiel ergibt sich:

$$(A.29)\quad E(x) = \sum_{i=1} x_i \, f(x_i)$$

$$= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3{,}5$$

Für kontinuierliche Zufallsvariablen wird die Summation durch eine Integration ersetzt:

$$(A.30)\qquad E(x) = \int_{-\infty}^{\infty} x \, f(x) \, dx$$

Schließlich tritt manchmal das Problem auf, daß ein Verteilungsgesetz nur für einen Teil der Grundgesamtheit gilt. Das Gesamtpopulationsmodell ergibt sich also durch Mischung der verschiedenen Verteilungen (*Mischverteilungen*). Auch dieses Problem löst man durch Integration (vgl. Gleichung 2.46 und 6.4).

| Verteilung | Exponentialverteilung | Weibull – Verteilung | Extremwertverteilung |
|---|---|---|---|
| Wertebereich der Zufallsvariablen | $t \geq 0$ | $t \geq 0$ | $-\infty < y < \infty$ |
| Parameter der Verteilung | $\lambda > 0$ | $\lambda > 0, \gamma > 0$ | $-\infty < \mu < \infty, \sigma > 0$ |
| Dichte f(t),f(y) | $\lambda \exp(-\lambda t)$ | $\lambda\gamma(\lambda t)^{\gamma-1} \exp(-(\lambda t)^{\gamma})$ | $\frac{1}{\sigma} \exp\left(\frac{y-\mu}{\sigma}\right) \exp\left[-\exp\left(\frac{y-\mu}{\sigma}\right)\right]$ |
| Überlebenswahrscheinlichkeit S(t), S(y) | $\exp(-\lambda t)$ | $\exp[-(\lambda t)^{\gamma}]$ | $\exp\left[-\exp\left(\frac{y-\mu}{\sigma}\right)\right]$ |
| Rate r(t), r(y) | $\lambda$ | $\lambda\gamma(\lambda t)^{\gamma-1}$ | $\frac{1}{\sigma} \exp\left(\frac{y-\mu}{\sigma}\right)$ |
| Mittelwert E(t), E(y) | $\lambda^{-1}$ | $\Gamma(1+\gamma^{-1})\lambda^{-1}$ | $\mu - E\sigma$ |
| Varianz VAR(t), VAR(y) | $\lambda^{-2}$ | $[\Gamma(1+2\gamma^{-1}) - \{\Gamma(1+\gamma^{-1})\}^2]\lambda^{-2}$ | $(\pi\sigma)^2 / 6$ |
| Verteilung für ln t / exp λ | Extremwertverteilung mit $\mu = -\ln\lambda$ und $\sigma = 1$ | Extremwertverteilung mit $\mu = -\ln_1\lambda$ und $\sigma = \gamma^{-1}$ | Weibull – Verteilung mit $\lambda = \exp(-\mu)$ und $\lambda = \sigma$ |
| Stochastisches Modell | zufällige Ereignisse (konstantes Risiko) | schwächstes Glied einer Kette | schwächstes Glied einer Kette |
| Vor – und Nachteile | Einfache Anwendung und Interpretation, wenig robust | Flexibilität, Test der Exponentialverteilung | Verbindung zur Weibull – Verteilung |
| Erweiterungen, Spezialfälle | – | Exponentialverteilung ($\gamma = 1$) Rayleigh – Verteilung ($\gamma = 2$) | Gompertz – Verteilung ($y \geq 0$) |
| Bemerkungen | – | $\Gamma(.)$ Gamma - Funktion | E = 0,5772 (Euler's Konstante) |

*Abbildung B.1: Exponentialverteilung*

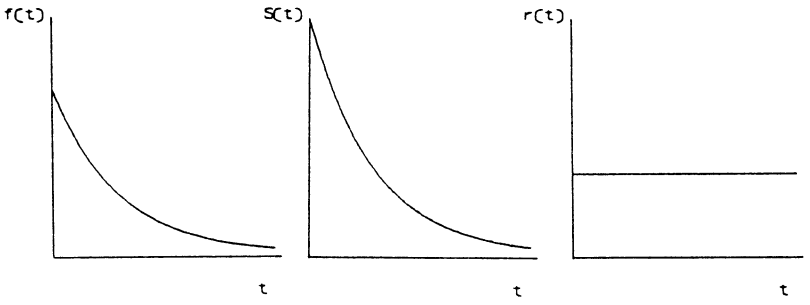
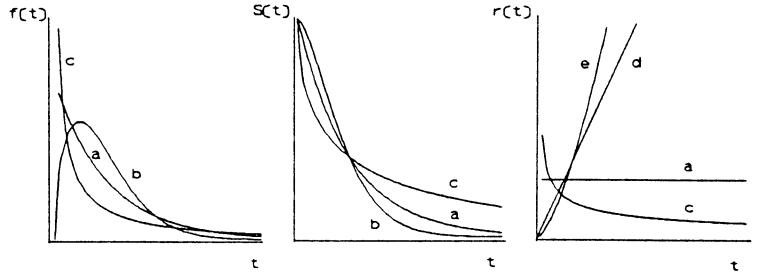
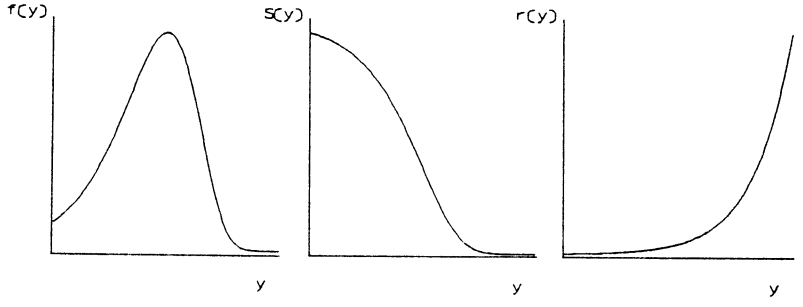



*Abbildung B.2: Weibull – Verteilung*



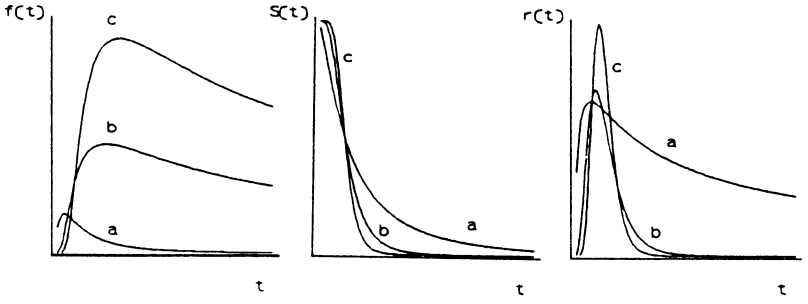

a) $\gamma = 1$, b) $\gamma > 1$, c) $\gamma < 1$, d) $\gamma = 2$, e) $\gamma > 2$

*Abbildung B.3: Standard – Extremwertverteilung*



| Verteilung | Lognormal – Verteilung | Normalverteilung | log – logistische Verteilung |
|---|---|---|---|
| Wertebereich der Zufallsvariablen | $t > 0$ | $-\infty < y < \infty$ | $t \geq 0$ |
| Parameter der Verteilung | $\lambda > 0, \gamma > 0$ | $-\infty < \mu < \infty, \sigma > 0$ | $\lambda > 0, \gamma > 0$ |
| Dichte | $\frac{\lambda}{t\sqrt{2\pi}} \exp\left(\frac{-[\gamma \ln(\lambda t)]^2}{2}\right)$ | $\frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right]$ | $\frac{\lambda\gamma(\lambda t)^{\gamma-1}}{(1+(\lambda t)^{\gamma})^2}$ |
| Überlebenswahrscheinlichkeit | $1 - \Phi[\gamma \ln(\lambda t)]$ | $1 - \Phi\left(\frac{y-\mu}{\sigma}\right)$ | $[1+(\lambda t)^{\gamma}]^{-1}$ |
| Rate | $\frac{\frac{\gamma}{t\sqrt{2\pi}} \exp\left(\frac{-[\gamma \ln(\lambda t)]^2}{2}\right)}{1 - \Phi[\gamma \ln(\lambda t)]}$ | $\frac{\frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right]}{1 - \Phi\left(\frac{y-\mu}{\sigma}\right)}$ | $\frac{\lambda\gamma(\lambda t)^{\gamma-1}}{1+(\lambda t)^{\gamma}}$ |
| Mittelwert | $\exp(0{,}5\,\gamma^{-2})\,\lambda^{-1}$ | $\mu$ | |
| Varianz | $[\exp(2\gamma^{-2}) - \exp(\gamma^{-2})]\,\lambda^{-2}$ | $\sigma^2$ | |
| Verteilung für $\ln t / \exp \lambda$ | Normalverteilung mit $\mu = -\ln_1\lambda$ und $\sigma = \gamma$ | Lognormalverteilung mit $\lambda = \exp(-\mu)$ und $\lambda = \sigma$ | logistische Verteilung mit $\mu = -\ln_1\lambda$ und $\sigma = \lambda$ |
| Stochastisches Modell | zunehmender Verschleiß aufgrund vieler voneinander unabhängiger Ursachen | zunehmender Verschleiß aufgrund vieler voneinander unabhängiger Ursachen | |
| Vor – und Nachteile | wechselndes Risiko, komplizierte Berechnung bei zensierten Daten | Verbindung der Lognormal – verteilung | Approximation zur Lognormal – verteilung |
| Erweiterungen | – | – | – |
| Bemerkungen | $\Phi(.)$ unvollständiges Integral der standardisierten Normalverteilung | $\Phi(.)$ unvollständiges Integral der standardisierten Normalverteilung | identisch mit Weibull – Verteilung außer Faktor $(1+(\lambda t)^{\gamma})^{-1}$ |

*Abbildung B.4: Log – Normalverteilung*




a) $\gamma = 1$, b) $\gamma = 2$, c) $\gamma = 3$

*Abbildung B.5: Standard – Normalverteilung*

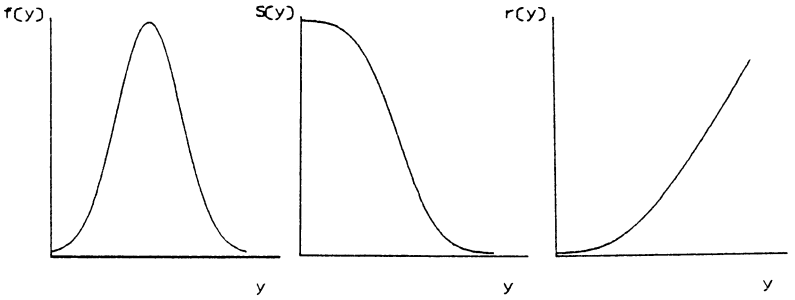
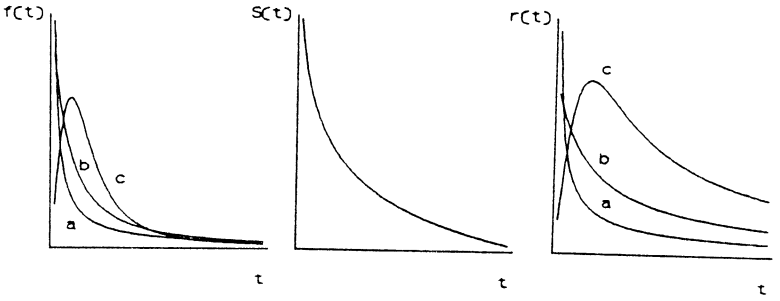



*Abbildung B.6: Log – logistische Verteilung*




a) $\gamma < 1$, b) $\gamma = 1$, $\gamma > 1$

*Anhang B: Eine Synopse verschiedener theoretischer Verteilungen III*

| Verteilung | Logistische Verteilung | Gompertz – Verteilung | Gamma – Verteilung |
|---|---|---|---|
| Wertebereich der Zufallsvariablen | $-\infty < y < \infty$ | $t \geq 0$ | $t > 0$ |
| Parameter der Verteilung | $-\infty < \mu < \infty, \sigma$ | $\lambda > 0, \gamma > 0$ ($\gamma < 0$) | $\lambda > 0, \gamma > 0$ |
| Dichte | $\frac{1}{\sigma}\left[1 + \exp\left(-\frac{y-\mu}{\sigma}\right)\right]^{-1}$ $\left[1 - \left\{1 + \exp\left(-\frac{y-\mu}{\sigma}\right)\right\}\right]^{-1}$ | $\lambda \exp(\gamma t) \exp\left[-\frac{\lambda}{\gamma}(\exp(\gamma t) - 1)\right]$ | $\lambda^{\gamma} t^{\gamma-1} \exp(-\lambda t) / \Gamma(\gamma)$ |
| Überlebenswahrscheinlichkeit | $1 - \left[1 + \exp\left(-\frac{y-\mu}{\sigma}\right)\right]^{-1}$ | $\exp\left[-\frac{\lambda}{\gamma}(\exp(\gamma t) - 1)\right]$ | $1 - I(\lambda t, \gamma)$ |
| Rate | $\frac{1}{\sigma}\left[1 + \exp\left(-\frac{y-\mu}{\sigma}\right)\right]^{-1}$ | $\lambda \exp(\gamma t)$ | $\frac{\lambda^{\gamma} t^{\gamma-1} \exp(-\lambda t)}{\Gamma(\gamma)\,[1 - I(\lambda t, \gamma)]}$ |
| Mittelwert | $\mu$ | $\lambda \exp\frac{\lambda}{\gamma} \int_{\lambda/\gamma}^{\infty} t^{-1} \exp(-t) dt$ | $\gamma / \lambda$ |
| Varianz | $(\pi\sigma)^2 / 3$ | – | $\gamma / \lambda^2$ |
| Verteilung für $\ln t / \exp \lambda$ | log – logistische Verteilung mit $\lambda' = \exp_1(-\mu)$ und $\gamma = \sigma$ | | Log – Gamma – Verteilung (vgl. LAWLESS 1982: 21ff.) |
| Stochastisches Modell | Wachstum von einem unteren zu einem oberen Schwellwert | Gompertz'sches Mortalitätsgesetz | Summe statistisch unabhängiger exponentiell verteilter Wartezeiten |
| Vor- und Nachteile | Verbindung zur log-logistischen Verteilung | einfache Anwendung, Personen ohne Ereignis ($\gamma < 0$) | komplizierte Berechnung, Test der Exponentialverteilung |
| Erweiterungen, Spezialfälle | – | Gompertz – Makcham – Verteilung $r(t) = \lambda_1 + \lambda_2 \exp(\gamma t)$ | Exponentialverteilung ($\gamma = 1$) $\chi^2$ – Verteilung ($\lambda = 0{,}5$, $\gamma = df/2$) |
| Bemerkungen | – | – | $\Gamma(.)$ Gamma – Funktion, $I(.,.)$ unvollständiges Gamma – Integral |

*Abbildung B.7: Logistische Verteilung*

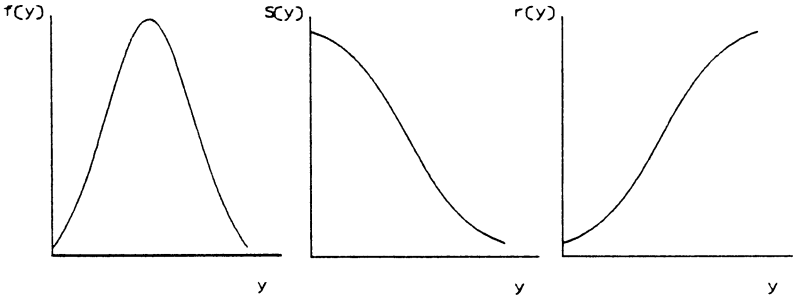
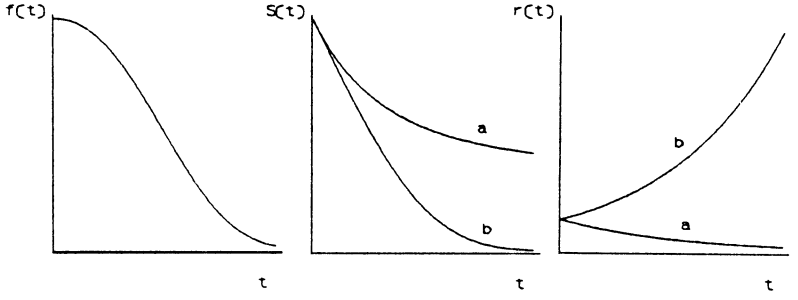
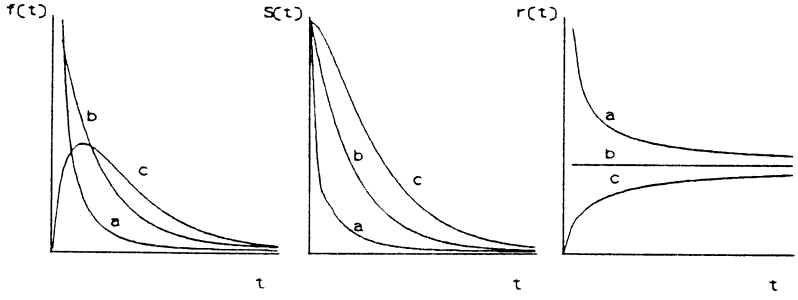



*Abbildung B.8: Gompertz – Verteilung*




a) $\gamma < 0$, b) $\gamma > 0$

*Abbildung B.9: Gamma – Verteilung*




a) $\gamma < 1$, b) $\gamma = 1$, c) $\gamma > 1$

# *Anhang C: Glossar*

## *Notation und Symbole*

| | |
|---|---|
| $x$ | Skalar |
| $\mathbf{x}$ | Vektor (fett, kleingeschrieben) |
| $\mathbf{X}$ | Matrix (fett, großgeschrieben) |
| $\mathbf{X}'$ | transponierte Matrix |
| $f'(x) = df(x)/dx$ | 1. Ableitung der Funktion f(x) nach x |
| $f''(x) = d^2f(x)/(dx)^2$ | 2. Ableitung der Funktion f(x) nach x |
| $\partial f(x,z)/\partial x$ | 1. partielle Ableitung der Funktion f(x,z) nach x |
| $\partial^2 f(x,z)/\partial x \partial x$ | 2. partielle Ableitung der Funktion f(x,z) nach x |
| $\int_0^x f(u)du$ | Integral von 0 bis x über f(x) |
| $u$ | Integrationsvariable |
| $\varkappa$ | Integrationskonstante |
| $E = 0{,}5772\ldots$ | Euler'sche Konstante |
| $\exp(\cdot)$ | Exponentialfunktion |
| $\ln(\cdot)$ | natürlicher Logarithmus |
| $\Gamma(.)$ | Gamma – Funktion |
| $g(\cdot)$ | Dichtefunktion, Verbindungsfunktion |
| $E(\cdot)$ | Erwartungswert |
| $VAR(\cdot)$ | Varianz |
| $COV(\cdot)$ | Kovarianz |
| $P(\cdot)$ | Wahrscheinlichkeit von |
| $\bar{x}$ | Arithmetisches Mittel der Variable x |
| $\ddot{x}$ | Median der Variable x |

## *Indizes, Exponenten*

| | |
|---|---|
| $i,l$ | Index für $\sum$ und $\prod$ |
| $i = 1,\ldots,N$ | Index Beobachtungen |
| $j = 1,\ldots,J$ | Index Ausgangszustände |
| $j = 1,\ldots,J$ | Index Gruppen |
| $k = 1,2,\ldots$ | Index der Realisationen $t_k$ |
| $k = 1,\ldots,K+1$ | Index Zeit – Intervalle |
| $k = 1,\ldots,K$ | Index Zielzustände |
| $\iota = 1,\ldots,I$ | Index Rangplätze |
| $p = 1,\ldots,P$ | Index Kovariate |
| $e = 1,\ldots,E_i$ | Index (hochgestellt) Episoden |
| $p = 1,\ldots,P$ | Index (hochgestellt) Perioden |
| $q = 1,\ldots,Q$ | Index (hochgestellt) Schichten |

## *Zufallsvariablen*

| | |
|---|---|
| T | Zufallsvariable (großgeschrieben) "Wartezeit" |
| t | Realisation (kleingeschrieben) von T |
| $t_k$ | Realisation von T (diskret) |
| Y | Zufallsvariable ln(T) |
| y | Realisation von ln(T) |
| Z(t) | Zufallsvariable "Zustand zum Zeitpunkt t" |
| $\mu$ | Lageparameter einer Verteilung |
| $\sigma$ | Streuungsparameter einer Verteilung |
| $\lambda$ | Skalierungsparameter einer Wartezeit – Verteilung |
| $\gamma$ | Formparameter einer Wartezeit – Verteilung |
| N(0,1) | Standard – Normalverteilung mit $\mu = 0$ und $\sigma^2 = 1$ |
| $\Phi(\cdot)$ | unvollständiges Integral der Standard – Normalverteilung |
| E(0,1) | Standard – Extremwertverteilung mit $\mu = 0$ und $\sigma^2 = 1$ |
| $G(1,\sigma^2)$ | Gamma – Verteilung mit $\mu = 1$ und Varianz $\sigma^2$ |
| $I(\cdot,\cdot)$ | unvollständiges Integral der Gamma – Verteilung |
| $Y \sim N(0,1)$ | Y ist standard – normalverteilt |
| $Y \approx N(0,1)$ | Y ist näherungsweise standard – normalverteilt |

## *Wartezeit – Verteilungen*

| | |
|---|---|
| $f(t_k)$ | Sterbewahrscheinlichkeit |
| $f(t)$, $f_j(t)$ | Dichte |
| $S(t)$, $S_j(t)$ | Überlebenswahrscheinlichkeit, – funktion |
| $F(t)$, $F_j(t)$ | Verteilungsfunktion |
| $r(t)$, $r_j(t)$ | (Hazard –) Rate |
| $H(t)$ | Integrierte Rate |
| $q_k$ | bedingte Sterbewahrscheinlichkeit (diskrete Rate) |
| $p_k$ | bedingte Überlebenswahrscheinlichkeit |
| $f^*_{jk}(t)$ | Sub – Dichte |
| $S^*_{jk}(t)$ | Pseudo – Überlebensfunktion |
| $F^*_{jk}(t)$ | Sub – Verteilungsfunktion |
| $r_{jk}(t)$ | (Übergangs –) Rate |
| $\pi_{jk}$ | unbedingte Übergangswahrscheinlichkeit von j nach k |
| $m_{jk}(t)$ | bedingte Übergangswahrscheinlichkeit von j nach k zum Zeitpunkt t |

## *Regressionsmodelle*

| | |
|---|---|
| $\beta_p$ | p – ter Regressionskoeffizient proportionales Risikomodell |
| $\alpha_p$ | p – ter Regressionskoeffizient Skalierungsmodell |
| $a_p$ | Antilogarithmus des p – ten Regressionskoeffizienten |
| $\boldsymbol{\beta}$ | $(P+1)\cdot 1$ – Vektor der Regressionskoeffizienten |
| $x_{ip}$ | p – te Kovariate für Beobachtung i |
| $x_{ip}(t)$ | p – te zeitabhängige Kovariate für Beobachtung i |
| $\mathbf{x}_i$ | $1 \cdot P$ – Vektor der Kovariaten für Beobachtung i |

| | |
|---|---|
| $G^{e-1}$ | Vorgeschichte der Episode e |
| $\epsilon$ | Fehlerterm |
| $\lambda_0(t)$ | Basis – Rate des proportionalen Risikomodells |
| $H_0(t)$ | Integrierte Basis – Rate des proportionalen Risikomodells |
| $S_0(t)$ | Basis – Überlebensfunktion des proportionalen Risikomodells |

## *Schätzverfahren*

| | |
|---|---|
| $\delta$ | Zensierungsindikator (1 = Ereignis, 0 = Zensierung) |
| $E_i$ | Gesamtzahl der Episoden der Beobachtung i |
| $L(\cdot)$, ln $L(\cdot)$ | Likelihood – Funktion, Log – Likelihood – Funktion |
| $PL(\cdot)$, ln $PL(\cdot)$ | partielle Likelihood – Funktion, partielle Log – likelihood – Funktion |
| $\Theta(\beta)$ | Score – Funktion |
| $\Psi$ | Hesse'sche Matrix |
| $I(\beta)$ | Informationsmatrix |
| $\hat{\beta}$ | Schätzwert des Parameters $\beta$ |
| $\tilde{\beta}$ | Schätzwert des Parameters $\beta$ (Alternativmethode A) |
| $\bar{\beta}$ | Schätzwert des Parameters $\beta$ (Alternativmethode B) |

## *Rangtests*

| | |
|---|---|
| $\varphi_{\iota j}$ | Rangziffer des Ereignisses auf Rangplatz $\iota$ in Gruppe j |
| $\phi_{\iota j}$ | Rangziffer der Zensierung auf Rangplatz $\iota$ in Gruppe j |
| $\omega_{\iota j}$ | Summe aller Rangziffern auf Rangplatz $\iota$ in Gruppe j |
| $\Omega_j$ | Summe aller Rangziffern der Gruppe j |
| $\psi_{\iota j}$ | Differenz beobachtete – erwartete Ereignisse (Rangplatz $\iota$, Gruppe j) |
| $\nu_{\iota j}$ | vgl. Varianz – Kovarianz – Matrix der Rangziffern |

## *Sonstiges*

| | |
|---|---|
| $t_0$ | Prozeßbeginn |
| $t_\iota$ | Wartezeit auf Rangplatz $\iota$ |
| $t^w, t^d, t^l$ | Wartezeit (withdrawn, dead, lost cases) |
| $\tau$ | extern gesetzte Zeitpunkte |
| $\tau_0, \tau_e, \tau_s$ | Untersuchungszeitraum (Beginn, Ende, Eintritt) |
| $\tau_{k-1}, \tau_k, \tau_{mk}, \Delta\tau_k$ | Intervallbeginn, – ende, – mitte, – breite |
| $[\tau_{k-1}, \tau_k)$ | Intervall von $\tau_{k-1}$ bis (aber nicht einschließlich) $\tau_k$ |
| $\Upsilon$ | Gesamtdauer aller Wartezeiten |
| U | Gesamtdauer unzensierter Wartezeiten |
| V | Gesamtdauer zensierter Wartezeiten |
| $N, N_0, N_1, N_2, N_3$ | (Sub –) Stichprobenumfang |
| $R$ | Risikomenge (kursiv) |
| n | Anzahl Fälle in der Risikomenge |
| $D$ | Ereignismenge (kursiv) |
| D | Anzahl der Ereignisse insgesamt |

| | |
|---|---|
| d | Anzahl der Ereignisse |
| $C$ | Menge (kursiv) der zensierten Beobachtungen |
| C | Anzahl der Zensierungen insgesamt |
| c | Anzahl der Zensierungen |
| w | Anzahl der ohne Ereignis ausgeschiedenen Fälle |
| l | Anzahl der nicht auffindbaren Fälle |

# *Anhang D: Programme zur Verlaufsdatenanalyse*

| Analysemethode | SAS Version 6 | BMDP Version 1990 | SPSSX SPSS/PC | TDA Version 3.3 | RATE Version 2 |
|---|---|---|---|---|---|
| **a) Beschreibung** | | | | | |
| Sterbetafel | ja | ja | ja | ja | |
| Kaplan – Meier | ja | ja | | | |
| verfügbare Plots | S(t),f(t),r(t) | S(t),f(t),r(t) | S(t),f(t),r(t) | | |
| graphische Tests | lnS(t),ln(−lnS(t)) | lnS(t), H(t) | lnS(t) | | |
| Ausgabe in Datei | ja | | ja | ja | |
| **b) Gruppenvergleiche** | | | | | |
| Tests für Gruppenvergleiche | Wilcoxon, Savage, LRatio | Wilcoxon, Savage, Tarone, Prentice | Wilcoxon | Wilcoxon, Savage, Tarone, Prentice | |
| Sonstige Tests | metrische Merkmale | Gruppentrend | | | |
| **c) Regressionsmodelle** | | | | | |
| 1. COX Regression | ja | ja | | ja | ja |
| partielle Tests | Wald | Wald, Score, LRatio | | | |
| zeitabh. Kovariate | ja | ja | | ja | |
| Schichtung | ja | ja | | ja | |
| Schätzung $S_0(t)$ | Prentice | Breslow | | Breslow | |
| schrittweise Regr. | ja | ja | | | |
| Behandlung von ties | Breslow, Efron, Exakt | Breslow | | Breslow | ignoriert |
| zeitdiskrete Regr. | ja | | | | |
| Ausgabe in Datei | ja | | | ja | |
| 2. Parametr. Modelle | | | | | |
| Zielvariable | ln(T) | ln(T) | | $r_{jk}(t)$ | $r_{jk}(t)$ |
| Verteilungsmodell | a – e | a – d | | a – j | a,f |
| Modell mit $e \sim G(1,\sigma^2)$ | | | | a – d,h | a |
| periodenspez. Mod. | Episodensplitting | Episodensplitting | | a | a,f |
| zeitabh. Kovariate | Episodensplitting | Episodensplitting | | ja | Episodensplitting |
| schrittweise Regr. | | ja | | | |
| Aargabe in Datei | ja | | | ja | |
| **d) Modellevaluation** | | | | | |
| Prognosen | ja | ja | | ja | |
| Annahmentests | ja | ja | | ja | |
| Residuenanalyse | ja | ja | | | |

Verteilungsmodelle: a) Exponential, b) Weibull, c) Lognormal, d) log – logistisch, e) Gamma, f) Gompertz – Makeham, g) Raleigh, h) Sichel, i) Inverse Gauss, j) Box – Cox

**Bestelladressen**

SAS SAS Institute,
GmbH Postfach 105340,
Neuenheimer Landstr. 28 – 30
6900 Heidelberg 1

Unterprogramme LIFETEST, LIFEREG, PHREG

BMDP Statistical Software Ltd.,
Cork Technology Park Model,
Farm Road Cork, Ireland

Unterprogramme 1L, 2L

SPSS SPSS Software,
Steinsdorfstr. 19,
8000 München 22

Unterprogramm SURVIVAL

RATE Zentrum für Umfragen, Methoden und Analysen e.V.,
B2, 1,
6800 Mannheim 1

TDA Götz Rohwer,
European University Institute CPN 2330,
50100 Firenze – Ferrovia, Italy,
Tel. 0039/55/ 5092 – 229 Fax 0039/ 55/5092 – 201,
EARN ROHWER @ IFIIUE.FI.CNR.IT

## *Anhang E: Daten – und Programmbeispiele*

Die Auswertungen in den Kapiteln 4 und 5 wurden mit den auf den folgenden Seiten abgedruckten Programmen durchgeführt. Es handelt sich um eine Auswahl von SAS – und TDA – Programmen – zwei Programmpaketen, die aus meiner Sicht besonders gut für die Verlaufsdatenanalyse geeignet sind. Eine vollständige Liste aller Programme (inkl. BMDP, LIMDEP, GLIM und SPSS – Beispielen) ist gegen Einsendung von zwei HD – Disketten und Freiumschlag bei mir erhältlich: Prof. Dr. H.J. Andreß, Universität Bielefeld, Fakultät für Soziologie, Postfach 8640, 4800 Bielefeld 1. Noch einfacher geht es über electronic mail: USOZF049@COMPAREX.HRZ.UNI – BIELEFELD.DE. Alle Beispiele verwenden die folgenden zwei Datensätze.

*a) MZ71 – Mikrozensus – Zusatzerhebung 1971*

Die Datei MZ71.DAT enthält die in Tabelle 3.3 genannten 14 Variablen vater, ausb, status, taetnr, ausgbesh, zielbesh, wechsl67, bundwehr, weitbild, tjob, tlabor, ts, tf und sf1. Die folgende Liste zeigt die ersten 10 Datensätze.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 157 | 13 | 96 | 1 | 59 | 157 | 0 | 0 | 0 | 9 | 0 | 0 | 9 | 2 |
| 60 | 10 | 60 | 1 | 100 | 100 | 0 | 0 | 0 | 9 | 0 | 0 | 9 | 0 |
| 220 | 8 | 14 | 1 | 62 | 100 | 0 | 0 | 0 | 3 | 0 | 0 | 3 | 2 |
| 39 | 8 | 38 | 1 | 100 | 116 | 0 | 0 | 0 | 2 | 0 | 0 | 2 | 2 |
| 59 | 10 | 60 | 1 | 100 | 116 | 1 | 0 | 0 | 6 | 0 | 0 | 6 | 1 |
| 283 | 10 | 60 | 1 | 100 | 137 | 1 | 0 | 0 | 5 | 0 | 0 | 5 | 3 |
| 60 | 10 | 60 | 1 | 100 | 100 | 0 | 0 | 0 | 11 | 0 | 0 | 11 | 0 |
| 60 | 10 | 60 | 1 | 100 | 100 | 0 | 0 | 0 | 9 | 0 | 0 | 9 | 0 |
| 60 | 10 | 60 | 1 | 98 | 98 | 0 | 0 | 0 | 11 | 0 | 0 | 11 | 0 |
| 43 | 10 | 59 | 1 | 112 | 112 | 0 | 0 | 0 | 10 | 0 | 0 | 10 | 0 |

*b) KFN – Strafentlassenendaten*

Die Datei KFN.DAT enthält die in Tabelle 3.1 genannten 11 Variablen und das Merkmal b_ende ("Wie wurde die Bewährungszeit beendet?"). Die insgesamt 12 Spalten der folgenden Liste entsprechen den Variablen person, haftgrnd, b_ende, b_zeit, s_monat, straftat, termine, schulden, arbeit, sozprobl, wohnung und wartzeit.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 19 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 19 |
| 2 | 1 | 0 | 37 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 37 |
| 3 | 1 | 3 | 19 | 2 | 1 | 1 | 0 | 1 | 1 | 1 | 2 |
| 4 | 0 | 2 | 27 | 2 | 1 | 1 | 1 | 1 | 1 | 0 | 2 |
| 5 | 1 | 3 | 7 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 7 |
| 6 | 1 | 1 | 26 | 10 | 1 | 0 | 0 | 0 | 0 | 0 | 10 |
| 7 | 0 | 1 | 31 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 31 |
| 8 | 1 | 0 | 37 | 10 | 1 | 1 | 0 | 0 | 0 | 0 | 10 |
| 9 | 1 | 3 | 18 | 6 | 1 | 0 | 1 | 0 | 0 | 0 | 6 |
| 10 | 0 | 3 | 21 | 5 | 1 | 0 | 1 | 0 | 0 | 0 | 5 |
| 11 | 0 | 0 | 37 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 37 |
| 12 | 1 | 0 | 37 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 37 |
| 13 | 0 | 0 | 37 | 12 | 1 | 0 | 0 | 0 | 0 | 0 | 12 |
| 14 | 0 | 1 | 32 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 32 |
| 15 | 0 | 1 | 33 | 16 | 1 | 0 | 1 | 1 | 0 | 0 | 16 |
| 16 | 0 | 0 | 37 | 7 | 1 | 0 | 0 | 0 | 0 | 0 | 7 |
| 17 | 0 | 3 | 17 | 4 | 1 | 0 | 1 | 0 | 0 | 0 | 4 |
| 18 | 0 | 2 | 20 | 12 | 1 | 0 | 1 | 0 | 1 | 0 | 12 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 19 | 1 | 0 | 37 | 21 | 1 | 0 | 1 | 0 | 0 | 0 | 21 |
| 20 | 0 | 1 | 35 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 35 |
| 21 | 1 | 3 | 28 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 28 |
| 22 | 1 | 0 | 37 | 16 | 1 | 0 | 0 | 0 | 0 | 0 | 16 |
| 23 | 1 | 1 | 33 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 33 |
| 24 | 0 | 3 | 23 | 2 | 1 | 1 | 1 | 0 | 1 | 0 | 2 |
| 25 | 1 | 2 | 7 | 7 | 1 | 1 | 1 | 1 | 1 | 0 | 7 |
| 26 | 1 | 1 | 37 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 37 |
| 27 | 0 | 1 | 25 | 14 | 1 | 0 | 0 | 0 | 0 | 0 | 14 |
| 28 | 0 | 1 | 34 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 34 |
| 29 | 1 | 1 | 33 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 33 |
| 30 | 0 | 2 | 7 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 3 |
| 31 | 0 | 3 | 14 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 3 |
| 32 | 1 | 1 | 33 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 33 |
| 33 | 1 | 2 | 18 | 16 | 1 | 0 | 0 | 1 | 0 | 0 | 16 |
| 34 | 1 | 1 | 8 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 8 |
| 35 | 1 | 1 | 36 | 12 | 1 | 0 | 1 | 0 | 0 | 0 | 12 |
| 36 | 0 | 2 | 10 | 2 | 1 | 1 | 0 | 1 | 1 | 0 | 2 |
| 37 | 1 | 3 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 |
| 38 | 0 | 1 | 35 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 35 |
| 39 | 0 | 2 | 16 | 3 | 1 | 0 | 0 | 0 | 1 | 0 | 3 |
| 40 | 0 | 3 | 12 | 3 | 1 | 0 | 0 | 0 | 1 | 0 | 3 |
| 41 | 1 | 3 | 21 | 14 | 1 | 0 | 0 | 0 | 1 | 0 | 14 |
| 42 | 0 | 1 | 36 | 16 | 1 | 1 | 0 | 1 | 1 | 0 | 16 |
| 43 | 0 | 0 | 35 | 3 | 1 | 0 | 0 | 0 | 1 | 0 | 3 |
| 44 | 1 | 2 | 2 | 4 | 1 | 1 | 1 | 1 | 1 | 1 | 4 |
| 45 | 0 | 1 | 27 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 27 |
| 46 | 0 | 0 | 35 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 35 |
| 47 | 1 | 0 | 34 | 14 | 1 | 0 | 1 | 1 | 1 | 0 | 14 |
| 48 | 0 | 2 | 8 | 6 | 1 | 0 | 0 | 0 | 0 | 0 | 6 |
| 49 | 1 | 1 | 22 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 22 |
| 50 | 1 | 1 | 10 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 10 |
| 51 | 1 | 0 | 33 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 33 |
| 52 | 1 | 3 | 15 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 15 |
| 53 | 1 | 0 | 28 | 17 | 1 | 0 | 1 | 0 | 0 | 0 | 17 |
| 54 | 1 | 3 | 8 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 8 |
| 55 | 0 | 0 | 27 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 27 |
| 56 | 0 | 2 | 7 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 7 |
| 57 | 0 | 0 | 27 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 27 |
| 58 | 0 | 3 | 20 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 20 |
| 59 | 0 | 0 | 24 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 24 |
| 60 | 0 | 1 | 23 | 14 | 1 | 0 | 1 | 0 | 0 | 0 | 14 |
| 61 | 0 | 0 | 24 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 24 |
| 62 | 1 | 0 | 23 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 23 |
| 63 | 1 | 0 | 16 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 16 |
| 64 | 0 | 3 | 11 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 11 |

In den großen sozialwissenschaftlichen Programmpaketen (SAS, BMDP, SPSS) werden die einzelnen Merkmale mit ihrem Variablennamen angesprochen. In den folgenden SAS−Programmen wird davon ausgegangen, daß bereits eine entsprechende SAS−Datei verlauf.mz71 bzw. verlauf.kfn mit den o.g. Akronymen eingerichtet wurde. Die beiden stand−alone−Programme RATE und TDA verwenden statt Namen die Position der jeweiligen Variablen auf dem Eingaberecord. Die Variable wartzeit der KFN−Daten befindet sich z.B. in der 12. Spalte der Datei und wird daher in TDA mit c12 (column 12) angesprochen. Die folgenden Programme sollen lediglich die prinzipiellen Möglichkeiten von SAS und TDA illustrieren und erheben daher keinen Anspruch auf Vollständigkeit. Sie sind ganz grob nach den Abschnitten dieser Arbeit gegliedert.

### *Abschnitt 4.2: Sterbetafelschätzung*

```
* ---------------------------------  PRG0401.SAS  --------------------------------- ;

OPTIONS PAGESIZE=65;

DATA tab3_2;
 t=_N_ - 1;
    INPUT abstieg          aufstieg              lateral              ohne;
   CARDS;
            0                 0                    0                    0
          129               223                   76                    0
          126               306                   92                    0
          170               354                   90                    0
          114               364                   70                    0
          119               368                   61                    0
           99               397                   49                    0
          125               409                   73                    0
          143               398                  130                    0
          115               251                  118                 1638
           65               172                  172                 1600
            0                 0                    0                 2050
 ;
RUN;

DATA tab4_1;
 SET tab3_2;
 KEEP t sf1 n;
 sf1=1; n= abstieg; output;
 sf1=2; n=aufstieg; output;
 sf1=3; n= lateral; output;
 sf1=0; n=    ohne; output;
RUN;

PROC LIFETEST DATA=tab4_1 METHOD=ACT WIDTH=1;
 FREQ n; TIME t*sf1(0,2,3);
RUN;
```

## *Abschnitt 4.3: Kaplan–Meier–Schätzung*

```
* --------------------------------  PRG0401.SAS  -------------------------------- ;

LIBNAME verlauf '.';          * betriebssystemspezifisch: MS/DOS;
OPTIONS PAGESIZE=65;

PROC LIFETEST DATA=verlauf.kfn METHOD=KM;
 TIME wartzeit*straftat(0);
RUN;
```

## *Abschnitt 4.4: Gruppenvergleiche*

```
* --------------------------------  PRG0401.SAS  -------------------------------- ;

OPTIONS PAGESIZE=65;

DATA tab3_8;
 t=_N_ - 1;
    INPUT  abstieg1          ohne1              abstieg2              ohne2;
   CARDS;
            0                  0                  0                    0
           19                 35                110                  264
           21                 39                105                  357
           35                 44                135                  398
           23                 21                 91                  413
           12                 35                107                  394
           24                 34                 75                  411
           26                 34                 99                  448
           25                 53                118                  475
           24                201                 91                 1779
           13                178                 52                 1728
            0                160                  0                 1829
;
RUN;

DATA tab3_8;
 SET tab3_8;
 sf1=1; n=abstieg1; qual=1; output;
 sf1=0; n=   ohne1; qual=1; output;
 sf1=1; n=abstieg2; qual=0; output;
 sf1=0; n=   ohne2; qual=0; output;
RUN;

PROC LIFETEST DATA=tab3_8 METHOD=ACT WIDTH=1 OUTS=plots;
 FREQ n; TIME t*sf1(0); STRATA qual;
RUN;

DATA plots;
 SET plots;
 IF midpoint GT 1;
 IF qual=1 THEN haz1=hazard  ; ELSE haz2=hazard ;
 IF qual=1 THEN ucl1=haz_ucl ; ELSE ucl2=haz_ucl;
 IF qual=1 THEN lcl1=haz_lcl ; ELSE lcl2=haz_lcl;
RUN;

GOPTIONS DEVICE=CGM GACCESS='SASGASTD>tab3_8.cgm'
         GSFMODE=REPLACE;
```

```
PROC GPLOT DATA=plots;
 PLOT (haz1 haz2 lcl1 ucl1)*midpoint/ OVERLAY;
 SYMBOL1 I=JOIN V=NONE L=1;
 SYMBOL2 I=JOIN V=NONE L=1;
 SYMBOL3 I=JOIN V=NONE L=33;
 SYMBOL4 I=JOIN V=NONE L=33;
RUN;


* ---------------------------------  PRG0401.SAS  --------------------------------- ;

LIBNAME verlauf '.';          * betriebssystemspezifisch: MS/DOS;
OPTIONS PAGESIZE=65;

PROC LIFETEST DATA=verlauf.kfn METHOD=KM;
 TIME wartzeit*straftat(0); STRATA sozprobl;
RUN;
```

## *Abschnitt 5.3/4: Parametrische Regressionsmodelle*

```
* ---------------------------------  PRG0401.SAS  --------------------------------- ;

df = c:\public\hja\verlauf\mz71.dat;    Datei mit Rohdaten
noc= 13075;                             13075 records
r  = 1-7587;                            Nur erste Taetigkeiten
ts = 0;                                 Beginn = 0 fuer alle
tf = c10;                               Ende = Taetigkeitsdauer
org= 0;                                 Ausgangszustand = 0 fuer alle
des= c14;                               Zielzustand

  v1(vater)   =c1;                      Kovariate:
  v2(ausb)    =c2;
  v3(status)  =c3;                      Aus Speicherplatzgruenden
* v4(taetnr)  =c4;                      sollte man die Variablen,
  v5(ausgbesh)=c5;                      die man nicht benoetigt,
  v6(zielbesh)=c6;                      zunaechst durch einen *
  v7(wechsl67)=c7;                      vom Einlesen ausschliessen.
* v8(bundwehr)=c8;
* v9(weitbild)=c9;
*v10(tjob)    =c10;
*v11(tlabor)  =c11;

* TDA kann in einem Lauf nur jeweils ein Modell schätzen.
* Waehlen Sie ein anderes Modell, indem Sie * loeschen.

  mod=2;                                Exponentialverteilung

  xa(0,1)= 0;                           Modell 0: Abstiege
  xa(0,2)= 0;                           Modell 0: Aufstiege
  xa(0,3)= 0;                           Modell 0: horiz. Mobilitaet

* mod=2;                                Exponentialverteilung

* xa(0,1)=1,2,3;                       Modell 1: Abstiege
* xa(0,2)=1,2,3;                       Modell 1: Aufstiege
* xa(0,3)=1,2,3;                       Modell 1: horiz. Mobilitaet
```

```
* mod=2;                                Exponentialverteilung

* xa(0,1)=1,2,3,5,6,7;                  Modell 2: Abstiege
* xa(0,2)=1,2,3,5,6,7;                  Modell 2: Aufstiege
* xa(0,3)=1,2,3,5,6,7;                  Modell 2: horiz. Mobilitaet

* mod=2;                                Exponentialverteilung

* xa(0,1)=1,2,3,5;                      Modell 3: Abstiege
* xa(0,2)=1,2,3,5;                      Modell 3: Aufstiege
* xa(0,3)=1,2,3,5;                      Modell 3: horiz. Mobilitaet

* mod=6;                                Gompertzverteilung

* xb(0,1)=1,2,3,5;                      Modell 4: Abstiege
* xb(0,2)=1,2,3,5;                      Modell 4: Aufstiege
* xb(0,3)=1,2,3,5;                      Modell 4: horiz. Mobilitaet

* mod=7;                                Weibullverteilung

* xa(0,1)=1,2,3,5;                      Modell 5: Abstiege
* xa(0,2)=1,2,3,5;                      Modell 5: Aufstiege
* xa(0,3)=1,2,3,5;                      Modell 5: horiz. Mobilitaet

* mod=3;                                Periodenspez. Expon.verteilung
* tp =  0(2)11;                         zweijaehrige Perioden

* xa(0,1)=1,2,3,5;                      Modell 6: Abstiege
* xa(0,2)=1,2,3,5;                      Modell 6: Aufstiege
* xa(0,3)=1,2,3,5;                      Modell 6: horiz. Mobilitaet


* ---------------------------------  PRG0401.SAS  --------------------------------- ;

LIBNAME verlauf '.';          * betriebssystemspezifisch: MS/DOS;
OPTIONS PAGESIZE=65;

PROC LIFEREG  DATA=verlauf.mz71(OBS=7587);
 TITLE 'Modell 3';
  Abstieg:MODEL tjob*sf1(0,2,3)=vater ausb status ausgbesh
                / D=EXPONENTIAL;
 Aufstieg:MODEL tjob*sf1(0,1,3)=vater ausb status ausgbesh
                / D=EXPONENTIAL;
  Lateral:MODEL tjob*sf1(0,1,2)=vater ausb status ausgbesh
                / D=EXPONENTIAL;
RUN;

PROC LIFEREG  DATA=verlauf.mz71(OBS=7587);
 TITLE 'Modell 5';
  Abstieg:MODEL tjob*sf1(0,2,3)=vater ausb status ausgbesh
                / D=WEIBULL;
 Aufstieg:MODEL tjob*sf1(0,1,3)=vater ausb status ausgbesh
                / D=WEIBULL;
  Lateral:MODEL tjob*sf1(0,1,2)=vater ausb status ausgbesh
                / D=WEIBULL;
RUN;
```

```
* ---------------------------------  PRG0401.SAS  --------------------------------- ;

LIBNAME verlauf '.';          * betriebssystemspezifisch: MS/DOS;
OPTIONS PAGESIZE=65;

DATA ctab5_4;
 INPUT sozprobl;
 CARDS;
 0
 1
 ;
RUN;

PROC PHREG DATA=verlauf.kfn;
 TITLE 'Modell 1';
 MODEL wartzeit*straftat(0)=sozprobl;
 BASELINE OUT=tab5_4 COVARIATES=ctab5_4 SURVIVAL=survival;
RUN;

PROC PRINT DATA=tab5_4;
RUN;

PROC PHREG DATA=verlauf.kfn;
 TITLE 'Modell 2';
 MODEL wartzeit*straftat(0)=termine schulden arbeit sozprobl wohnung;
RUN;

PROC PHREG DATA=verlauf.kfn;
 TITLE 'Modell 3';
 MODEL wartzeit*straftat(0)=termine schulden arbeit wohnung;
 STRATA sozprobl;
 BASELINE OUT=prognose LOGLOGS=loglogs;
RUN;

AXIS1 LABEL=('ln(-lnS(t))');
SYMBOL1 I=JOIN V=NONE L=1;
SYMBOL2 I=JOIN V=NONE L=2;

PROC GPLOT DATA=prognose;
 PLOT loglogs*wartzeit=sozprobl/ VAXIS=AXIS1;
RUN;

PROC PHREG DATA=verlauf.kfn;
 TITLE 'Modell 4';
 MODEL wartzeit*straftat(0)=sozprobl t_sozpro;
 t_sozpro=sozprobl*(log(wartzeit)-log(11.27));
RUN;


* ---------------------------------  PRG0401.SAS  --------------------------------- ;

df = c:\daten\verlauf\kfn.dat;          Datei mit Rohdaten
ts = 0;                                 Beginn = 0 fuer alle
tf = c12;                               Ende = Wartezeit
org= 0;                                 Ausgangszustand = 0 fuer alle
des= c6;                                Zielzustand = Straffaellig?
nosort;
```

```
v1(termine) =c7;                        Kovariate
v2(schulden)=c8;
v3(arbeit)  =c9;
v4(sozprobl)=c10;
v5(wohnung) =c11;

  mod=1;                                Cox Regression

* TDA kann in einem Lauf nur jeweils ein Modell schätzen.
* Waehlen Sie ein anderes Modell, indem Sie * loeschen.

  xa(0,1)= 4;                           Modell 1: nur sozprobl
  prs=c:\daten\verlauf\basekfn.dat;     Datei für S(t)-Schaetzer
  pv4=0,1;                              Kovariatenkonstellationen

* xa(0,1)= 1,2,3,4,5;                  Modell 2: alle Kovariaten

* xa(0,1)= 1,2,3,  5;                  Modell 3: Schichtung
* grp(ohne)=c10[0];                    Schicht 1: ohne soziale P.
* grp(mit )=c10[1];                    Schicht 2: mit sozialen P.

* v6(t.sozpro)=v4*(log(TIME)-2.422);
* xa(0,1)= 4,6;                        Modell 4: Proportionalitaet
```

## *Abschnitt 5.6.3: Regressionsmodelle mit unbeobachteter Heterogenität*

```
* --------------------------------  PRG0401.SAS  -------------------------------- ;

df = c:\daten\verlauf\kfn.dat;          Datei mit Rohdaten
ts = 0;                                 Beginn = 0 fuer alle
tf = c12;                               Ende = Wartezeit
org= 0;                                 Ausgangszustand = 0 fuer alle
des= c6;                                Zielzustand = Straffaellig?

v1(termine) =c7;                        Kovariate
v2(schulden)=c8;
v3(arbeit)  =c9;
v4(sozprobl)=c10;
v5(wohnung) =c11;

  mod=2;                                Exponentialverteilung
  mix=1;                                Gamma-Mischverteilung

* TDA kann in einem Lauf nur jeweils ein Modell schätzen.
* Waehlen Sie ein anderes Modell, indem Sie * loeschen.

  xa(0,1)= 0;                           Modell 0: keine Kovariaten
* xa(0,1)= 4;                           Modell 1: nur sozprobl
* xa(0,1)= 1,2,3,4,5;                  Modell 2: alle Kovariaten
```

# *Verzeichnis der Tabellen und Abbildungen*

## *Abbildungen*

## *Tabellen*

# *Literaturverzeichnis*


Abramowitz, M./Stegun, I.A. (1965): Handbook of mathematical functions. New York: Dover

Aitkin, M.A./Anderson, D./Francis, B.J./Hinde, J.P. (1989): Statistical modelling in GLIM. Oxford: Clarendon Press

Allison, P.D. (1982): Discrete-time methods for the analysis of event histories. in: Sociological Methodology 1982 (Leinardt, S., ed.). San Francisco: Jossey Bass. 61-98

Allison, P.D. (1984): Event history analysis. Regression for longitudinal data. Beverly Hill/London: Sage

Andreß, H.J. (1980): Methoden temporaler Analyse. Bielefeld: Arbeitsberichte und Forschungsmaterialien Nr. 9 der Fakultät für Soziologie der Universität Bielefeld

Andreß, H.J. (1982): Tätigkeitswechsel und Berufserfahrung - Analyse zeitbezogener Daten mit Hilfe von Sterbetafeln an Hand eines Beipiels aus der Mobilitätsforschung. ZfS 4: 380-400

Andreß, H.J. (1983): The first ten years of a working career: an illustration of event-history analysis with West German mobility data. Computational Statistics and Data Analysis 1: 111-135

Andreß, H.J. (1984a): Determinanten der Rückfälligkeit ehemaliger Straffälliger - Analyse zeitbezogener Daten in der Kriminologie. in: Methodologische Probleme in der kriminologischen Forschungspraxis (Kury, H., ed.). Köln: Carl Heymanns. 421-452

Andreß, H.J. (1984b): Die ersten 10 Berufsjahre - Methodische Probleme der Analyse von Längsschnittdaten an Hand eines empirischen Beispiels aus der Mobilitätsforschung. Beiträge zur Arbeitsmarkt- und Berufsforschung 87. Nürnberg: Bundesanstalt für Arbeit

Andreß, H.J. (1985): Lineare Modelle der Arbeitslosigkeitsdauer. Analyse gruppierter Zeitdauern mit Hilfe der Minimum-Chi-Quadrat-Methode (GSK-Ansatz). Allgemeines Statistisches Archiv 69: 337-361

Andreß, H.J. (1989): Recurrent unemployment - the West-German experience. An application of count data models to panel data. European Sociological Review 5: 275-297

Andreß, H.J./Best, H./Sombert, K. (1992): Die Mandatsdauer im deutschen Reichstag 1867–1918: Eine Anwendung neuer Methoden der Analyse historischer Verlaufsdaten. Erscheint in: Neue Methoden der Analyse historischer Daten (Best, H./Thome, H., eds.). Historisch–Sozialwissenschaftliche Forschungen Bd. 23.

Arminger, G. (1984): Modelltheoretische und methodische Probleme bei der Analyse von Paneldaten mit qualitativen Variablen. Vierteljahreshefte des DIW 4: 470–480

Arminger, G./Müller, F. (1990): Lineare Modelle zur Analyse von Paneldaten. Opladen: Westdeutschem Verlag

Barlow, R.E./Proschan, F. (1965): Mathematical theory of reliability. New York: Wiley

Becker, R. (1990): Arbeitsmärkte im öffentlichen Dienst und in der Privatwirtschaft. Eine Längsschnittuntersuchung aus der Perspektive von Berufsverläufen. Zeitschrift für Soziologie 19: 360–375

Berkson, J./Gage, R. (1950): Calculation of survival rates for cancer Proceedings of the Mayo Clinic 25: 270ff.

Blossfeld, H.–P. (1989): Kohortendifferenzierung und Karriereprozeß. Frankfurt: Campus

Blossfeld, H.–P./Hamerle, A./Mayer, K.U. (1986): Ereignisanalyse. Frankfurt/M.: Campus

Brenner, M.H. (1979): Wirtschaftskrisen, Arbeitslosigkeit und psychische Erkrankungen. München: Urban & Schwarzenberg

Breslow, N.E. (1970): A generalized Kruskal–Wallis test for comparing k–samples subject to unequal patterns of censorship. Biometrika 57: 579–594

Breslow, N.E. (1974): Covariance analysis of censored data. Biometrika 30: 89–99

Buss, A. (1974): A general developmental model for interindividual differences, intraindividual differences, and intraindividual changes. Develop. Psych. 10 (1): 70–78

Carroll, G.R./Hannan, M.T./Tuma, N.B./Warsavage, B. (1978): The impact of measurement error in the analysis of log–linear rate models: monte carlo findings. Stanford, CA: Stanford University, Laboratory for Social Research, Techn. Report No. 69

Carroll, G.R./Mayer, K.U. (1984): Organizational effects in the wage attainment process. Social Science Journal 3: 5–22

Cattell, R.B. (1946): Description and measurement of personality. Yonkers, NY: World Book

Cattell, R.B. (1952): The three basis factor–analytic research designs: their interrelations and derivatives. Psych. Bull. 49: 499–520

Chamberlain, G. (1984): Panel data. in: Handbook of econometrics (Griliches, Z./Intriligator, M.D., eds.). Amsterdam/New York: Elsevier. 1247–1318

Chambers, J.M. (1977): Computational methods for data analysis. New York: Wiley & Sons

Chiang, C.L. (1968): Introduction to stochastic processes in biostatistics. New York: Wiley

Coleman, J.S. (1968): The mathematical study of change. in: Methodology in social research (H.M.Blalock Jr./Blalock,A., eds.). New York: McGraw–Hill

Coleman, J.S. (1981): Longitudinal data analysis of attribute data. New York: Basic Books

Cox, D.R. (1972): Regression models and life tables. J. R. Stat. Soc., B, 34: 187–220

Cox, D.R. (1975): Partial likelihood. Biometrika 62: 269–276

Cox, D.R./Oakes, D. (1984): Analysis of survival data. London: Chapman and Hall

Crowley, J./Hu, M. (1977): Covariance analysis of heart transplant survival data. Journal of the American Statistical Association 72: 27–36

Crowley, J./Storer, B.E. (1983): Comment on "A reanalysis of the Stanford heart transplant data". Journal of the American Statistical Association 78: 277–281

Cutler, S.F./Ederer, F. (1958): Maximum utilization of the life table method in analyzing survival. J. Chron. Diseases 8: 699–713

Davis, D.J. (1952): An analysis of some failure data. J. Am. Stat. Ass. 47: 113–150

Diekmann, A. (1984): Zur Analyse von Zeitintervallen unter Berücksichtigung unbeobachteter Heterogenität. Anwendungen auf Rückfallintervalle nach der Haftentlassung und die Dauer der Arbeitslosigkeit. Wien: Institut für Höhere Studien, Forschungsbericht Nr.197

Diekmann, A. (1990): Der Einfluß schulischer Bildung und die Auswirkungen der Bildungsexpansion auf das Heiratsverhalten. Zeitschrift für Soziologie 19: 265–277

Diekmann, A. (1990): Hazard rate models of social diffusion processes. Paper presented on the meeting of the American Sociological Association, Washington, August 11–15th 1990

Diekmann, A./Mitter, P. (1983): The `sickle–hypothesis`: a time dependent poisson model with applications to deviant behaviour and occupational mobility. Journal of Mathematical Sociology 9: 85–101

Diekmann, A./Mitter, P. (1984a): Methoden zur Analyse von Zeitverläufen: Anwendungen stochastischer Prozesse bei der Untersuchung von Ereignisdaten. Stuttgart: Teubner

Diekmann, A./Mitter, P. (1984b): A comparison of the `sickle function` with alternative stochastic models ofdivorce rates for Austrian and U.S. marriage cohorts. in: Progress in stochastic modelling of social processes (Diekmann,A./Mitter,P. eds.). New York: Academic Press

Diekmann, A./Preisendörfer, P. (1988): Turnover and employment stability in a large West German company. European Sociological Review 4: 233–248

Diprete, T.A. (1981): Employment and unemployment over the life–cycle. Racial differences and the effect of changing economic conditions. AJS 2: 286–307

Draper, N.R./Smith, H. jr. (1981): Applied regression analysis. 2nd ed. New York: Wiley

Dyer, A.R. (1975): An analysis of relationship of systolic blood pressure, serum cholesterol and smoking to 14–year mortality in the Chicago Peoples Gas Company study. Journal of Chronic Diseases 28: 565–570

Efron, B. (1977): Efficiency of Cox's likelihood function for censored data. J. Am. Stat. Assoc. 72: 557–565

Elandt–Johnson, R.C./Johnson, N.L. (1980): Survival models and data analysis. New York: Wiley

Feller, W. (1971): Introduction to probability theory and its applications. Vol. 2. New York: Wiley

Fennell, M.L./Tuma, N.B./Hannan, M.T. (1977): Quality of maximum likelihood estimates of parameters in a log−linear ratemodel. Stanford, CA: Stanford University, Laboratory for Social Research, Technical Report No. 59

Flinn, C.J./Heckman, J.J. (1982): Models for the analysis of labor force dynamics. in: Advances in econometrics (Rhodes, G./Basmenn, R., eds.). London, CT: JAI Press. 35−95

Forthofer, R.N./Lehnen, R.G. (1981): Public program analysis: a new categorical data approach. Belmont, CA: Wadsworth

Freeman, J./Carroll, G.R./Hannan, M.T. (1983): The liability of newness: Age dependence in organizational death rates. Am. Sociol. Rev. 48: 692−710

Galler, H.P. (1986): Übergangsratenmodelle bei intervalldatierten Ereignissen. Statistische Hefte 27: 1−22

Galler, H.P. (1988): Ratenmodelle mit stochastisch abhängigen konkurrierenden Risiken. Frankfurt/Mannheim: SFB 3 Mikroanalytische Grundlagen der Gesellschaftspolitik, Arbeitspapier Nr. 261

Gehan, E.A. (1965): A generalized Wilcoxon test for comparing arbitrarily singly censored samples. Biometrika 52: 203−223

Gehan, E.A./Siddiqui, M.M. (1973): Simple regression methods for survival time studies. Journal of the American Statistical Association 68: 848−856

Glenn, N.D. (1977): Cohort analysis. London/Beverly Hills: Sage Pubns.

Gompertz, B. (1825): On the nature of the function expressive of the law of human mortality. Phil. Trans. R. Soc. (London) 115: 513−583

Goodman, L.A./Magidson, J. (eds.) (1978): Analyzing qualitative, categorical data: log−linear models and latent structure analysis. Cambridge, Mass.: Abt

Grizzle, J.E./Starmer, C.F./Koch, G.G. (1969): Analysis of categorical data by linear models. Biometrics 25: 489−504

Gross, A.J./Clark, V. (1975): Survival distributions: reliability applications in the biomedical sciences. New York: Wiley

Gumpel, E.J. (1958): Statistics of extremes. New York: Columbia University Press

Hagenaars, J.A. (1990): Categorical longitudinal data. Log−linear panel, trend, and cohort analysis. London/Beverly Hills: Sage

Hajek, J. (1969): A course in nonparametric statistics. San Francisco: Holden−Day

Hajek, J./Sidak, Z. (1967): Theory of rank tests. New York: Academic Press

Hamerle, A. (1989): Multiple−spell regression models for duration data. Applied Statistics 38: 127−138

Hamerle, A./Tutz, G. (1989): Diskrete Modelle zur Analyse von Verweildauer und Lebenszeit. Frankfurt: Campus

Handl, J. (1977): Sozio−ökonomischer Status und der Prozess der Statuszuweisung − Entwicklung und Anwendung einer Skala. in: Klassenlagen und Sozialstruktur (Handl, J./Mayer, K.U./Müller, W., eds.). Frankfurt/Main: Campus

Hanefeld, U. (1984): Das sozio−ökonomische Panel. Eine Längsschnittstudie für die Bundesrepublik Deutschland. Vierteljahreshefte des DIW 4: 391−406

Hannan, M.T./Carroll, G.R. (1981): Dynamics of formal political structure: an event history analysis. American Sociological Review 46: 19−35

Hannan, M.T./Tuma, N.B./Groeneveld, L.P (1977): Income and marital events: Evidence from an income maintenance experiment. Am. J. Sociol. 82: 1186−1211

Hanushek, E.A./Jackson, J.E. (1977): Statistical methods for social scientists. New York: Academic Press

Hartung, J. (1982): Statistik − Lehr− und Handbuch der angewandten Statistik. München: Oldenbourg

Heckman, J.J. (1979): Sample selection bias as a specification error. Econometrica 47: 153−161

Heckman, J.J. (1981): Incidental parameters problem and the problem of initial conditions in estimating a discrete time − discrete data stochastic process and some monte carlo evidence. in: Structural analysis of discrete data: With econometric applications (Manski, C./McFadden, D., eds.). Cambridge, MASS.: MIT Press

Heckman, J.J./Singer, B. (1984): A method for minimizing the impact of distributional assumptions in econometric models for duration data. Econometrica 52: 271–320

Holt, J.D. (1978): Competing risk analysis with special reference to matched pair experiments. Biometrika 65: 159–166

Hsiao, C. (1986): Analysis of panel data. Cambridge: Cambridge University Press

Johnson, N.L./Kotz, S. (1970): Continuous univariate distributions, 2 vols. Boston, Mass.: Houghton Mifflin

Johnson, W.D./Koch, G.G. (1978): Linear models analysis of competing risks for grouped survival times. International Statistical Review 46: 21–51

Kalbfleisch, J.D./Prentice, R.L. (1973): Marginal likehoods based on Cox's regression and life model. Biometrika 64: 47–50

Kalbfleisch, J.D./Prentice, R.L. (1980): The statistical analysis of failure time data. New York: Wiley

Kaplan, E.L./Meier, P. (1958): Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc. 53: 457–481

Kessler, R.G./Greenberg, D.F. (1981): Linear panel analysis. Models of quantitative change. New York: Academic Press

King, G. (1989): Unifying political methodology: The likelihood theory of statistical inference. New York: Cambridge Univ. Press

Koch, G.G./Johnson, W.D./Tolley, H.D. (1972): A linear models approach to the analysis of survival and extent of disease in multidimensional contingency tables. Journal of the American Statistical Association 67: 783–796

Lagakos, S.W. (1980): The graphical evaluation of explanatory variables in proportional hazards regression models. Biometrika 68: 93–98

Lancaster, T. (1990): The econometric analysis of transition data. Cambridge: University Press

Lawless, J.F. (1982): Statistical models and methods for lifetime data. New York: Wiley

Lee, E.T. (1980): Statistical methods for survival data analysis. Belmont, CA: Wadsworth

Lee, E./Desu, M. (1972): A computer program for comparing samples with right censored data. Comp. Progr. Biomed. 2: 315–321

Lehmann, E.L. (1975): Nonparametrics: statistical methods based on ranks. San Francisco: Holden–Day

Littell, A.S. (1952): Estimation of the t–year survival rate from follow–up studies over a limited period of time. Human Biology 24: 87–116

Makeham, W.M. (1860): On the law of mortality and the construction of annuity tables. J. Inst. Actuaries (London) 18: 317–322

Mann, N.R./Schafer, R.E./Singpurwalla, N.D. (1974): Methods for statistical analysis of reliability and lifetime data. New York: Wiley

Mantel, N. (1966): Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chem. Ther. Rep. 50: 163–170

Miller, R.G. (1976): Least squares regression with censored data. Biometrika 63: 449–464

Miller, R.G. (1981): Survival analysis. New York: Wiley

Murray, W. (ed.) (1972): Numerical methods for unconstrained optimization. London/New York

Müller, W. (1977): Schulbildung und Weiterbildung als soziologische Hintergrundsvariablen. in: Sozialstrukturanalyse mit Umfragedaten (Pappi, F.U., ed.). Kronberg: Athenaeum

Müller, W. (1979): Muster beruflicher Karrieren in der Bundesrepublik Deutschland. Papier für die Konferenz des Research Committee Social Stratification der International Sociological Association, Berlin, 30.10.–2.11.1979

Namboodiri, K./Suchindran, C.M. (1987): Life table techniques and their applications. New York: Academic

Nelson, W. (1972): Theory and applications of hazard plotting for censored failure data. Technometrics 14: 945–966

Nelson, W. (1982): Applied life data analysis. New York: Wiley

Oakes, D. (1977): The assymptotic information in censored survival data. Biometrika 64: 441–448

Peto, R. (1972): Discussion of paper by D.R. Cox. J. R. Stat. Soc., B, 34: 205–207

Peto, R./Peto, J. (1972): Asymptotically efficient rank invariant procedures (with discussion). J. R. Stat. Soc., A, 135: 185–206

Plewis, I. (1985): Analysing change. Measurement and explanation using longitudinal data. Chichester/New York/Brisbane/Toronto/Singapore: Wiley

Plotnick, R. (1983): Turnover in the AFDC population: an event history analysis. Journal of Human Resources 18: 65–81

Pötter, U. (1989): Kleinstquadratschätzer bei zensierten Daten. Frankfurt/Mannheim: Sfb 3 Mikroanalytische Grundlagen der Gesellschaftspolitik, Arbeitspapier Nr. 305

Prentice, R.L. (1978): Linear rank tests with right–censored data. Biometrika 65: 167–179

Rao, C.R. (1973): Linear statistical inference and its applications. New York: Wiley

Rohwer, G. (1991): Analysis of transition data. A practical introduction with RATC. mimeo

Schneider, H. (1988): Determinanten der Arbeitslosigkeitsdauer. Eine mikroökonometrische Analyse auf der Grundlage von Längsschnittdaten des Sozio–ökonomischen Panels. Frankfurt/New York: Campus

Schneider, H. (1991): Verweildaueranalyse mit GAUSS. Frankfurt/New York: Campus

Singer, B./Spilerman, S. (1974): Social mobility models for heterogenous populations. in: Sociological Methodology 1973–1974 (H.L. Costner, ed.). San Francisco:Jossey–Bass, 256ff.

Singer, B./Spilerman, S. (1976): Some methodological issues in the analysis of longitudinal surveys. Ann. Econ. Soc. Meas. 5: 447–474

Singer, B./Spilerman, S. (1976): The representation of social processes by Markov models. Am. J. Sociol. 82: 1–54

Sörensen, A.B. (1977): Estimating rates from retrospective questions. in: Sociological Methodology 1977 (D.R. Heise, ed.). San Francisco: Jossey–Bass, 209–223

Sörensen, A.B./Tuma, N.B. (1978): Labor market structures and job mobility. Madison, WIS: Univ. Wisconsin, Working Paper No. 505–78

Sörensen, A./Sörensen, A.B. (1983): Modeling interdependence of life course events with event–history data. Paper prep. for the general session of the section for methodology at the meetings of the American Sociological Association in Detroit, Mich., September 2, 1983

Stange, K. (1970): Angewandte Statistik. Teil 1: Eindimensionale Probleme. Berlin: Springer

Tarone, R.E./Ware, J. (1977): On distribution–free tests for equality of survival distributions. Biometrika 64: 156–160

Tegtmeyer, H. (1976): Berufliche und soziale Umschichtung der Bevölkerung. Zeitschrift für Bevölkerungswissenschaft 1: 4–33

Tsiatis, A. (1978): A nonidentifiability aspect of the problem of competing risks. Proceedings of the National Academy of Sciences 72: 20–22

Tuma, N.B. (1976): Rewards, ressources and rate of mobility: a nonstationary multivariate stochastic model. Am. Sociol. Rev. 41: 338–360

Tuma, N.B. (1980): Invoking rate. Program manual. Stanford, CA: Stanford University

Tuma, N.B. (1982): Nonparametric and partially parametric approaches to event–history analysis. in: Sociological Methodology 1982 (Leinhardt, S., ed.). San Francisco: Jossey–Bass. 1–60

Tuma, N.B. (1982): Effects of labor market structure on job–shift patterns. in: The analysis of longitudinal labor market data (Heckman,J./ Singer, B., eds.). New York: Academic Press

Tuma, N.B./Hannan, M.T. (1978): Approaches to the censoring problem in analysis of event histories. in: Sociological Methodology 1979 (K. Schuessler, ed.). San Francisco: Jossey–Bass

Tuma, N.B./Hannan, M.T. (1984): Social dynamics: models and methods. Orlando: Academic Press

Tuma, N.B./Hannan, M.T./Groeneveld, L.P. (1979): Dynamic analysis of event histories. Am. J. Sociol., 84: 820–854

Walsh, G.R. (1975): Methods of optimization. New York: Wiley

Weibull, W. (1951): A statistical distribution function of wide applicability. J. Appl. Mech. 18: 293–297

Ziegler, R./Brüderl, J./Diekmann, A. (1988): Stellensuchdauer und Anfangseinkommen bei Hochschulabsolventen. Ein empirischer Beitrag zur Job-Search-Theorie. Zeitschrift für Wirtschafts- und Sozialwissenschaften 108: 247-270